REPORT DOCUMENTATION PAGE

AFRL-SR-AR-TR-04-

0220

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
31-03-04	Final Performance Report	02-15-03 - 12/31/03

4. TITLE AND SUBTITLE
THE ROLE OF INDIVIDUAL AND TEAM COGNITION IN UNINHABITED AIR VEHICLE COMMAND-AND-CONTROL

5a. CONTRACT NUMBER

5b. GRANT NUMBER
F49620-03-1-0024

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Nancy J. Cooke, Janie A. DeJoode, Harry K. Pedersen, Jamie C. Gorman, Olena O. Connor, Preston A. Kiekel

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Arizona State University East
7001 E. Williams Field Rd
Mesa, AZ 85212

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Dr. Robert Sorkin
AFOSR/NL
4015 Wilson Blvd.
Arlington, VA 22203-1954

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report documents a three-year AFOSR-funded research effort designed to study individual and team cognition in Unmanned Aerial Vehicle command-and-control. Three experiments were conducted in the CERTT lab's UAV-STE (Unmanned Aerial Vehicle – Synthetic Task Environment). Experiments 1 and 2 had two main manipulations, dispersion (co-located vs. distributed) and workload (low or high) and consisted of 20 teams flying multiple 40-minute missions. The results from these experiments indicate that team performance was affected by increased workload, but not impacted by the dispersion condition, although dispersion did affect knowledge and team process.

In Experiment 3 data were collected in CERTT's UAV-STE from five expert teams in order to benchmark team performance. The task acquisition curve was accelerated for several of these teams. What differentiated the expert teams were their long histories of working together in a networked setting (e.g., internet video games). It appears that this background alone was enough to speed up their task acquisition in terms of both team process and performance.

In addition to these three experiments, this report also documents archival analyses in which we identify individual and role-specific characteristics that are associated with team performance and find support for the utility of holistic and on-line knowledge elicitation in order to accurately assess team knowledge as it relates to team performance. The findings of this report can be summarized broadly by the implication for an increased focus on ongoing, coordinative and other process behaviors rather than focusing on static knowledge for improving applications, theory, and methodology, as they relate to team cognition.

15. SUBJECT TERMS
Team performance, team cognition, distributed mission environments, unmanned aerial vehicles

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	UL		Nancy J. Cooke
					19b. TELEPHONE NUMBER (include area code) 480-727-1331

THE ROLE OF INDIVIDUAL AND TEAM COGNITION IN UNINHABITED AIR VEHICLE COMMAND-AND-CONTROL

Nancy J. Cooke, Janie A. DeJoode, Harry K. Pedersen, Jamie C. Gorman, Olena O. Connor, and Preston A. Kiekel

FINAL PERFORMANCE REPORT

31 March 2004

AFOSR Grants F49620-01-1-0261and F49620-03-1-0024

Performance Period: February 2001 – December 2003

Contact Information

Nancy J. Cooke, Ph.D.
Applied Psychology Program
Arizona State University East
7001 E. Williams Field Rd. Bldg. 140
Mesa, AZ 85212

Email: ncooke@asu.edu
Web Sites: www.certt.com
www.cerici.org
Phone: 480-727-1331
Fax: 480-727-1363

TABLE OF CONTENTS

List of Figures

List of Tables

List of Appendices

1.0 EXECUTIVE SUMMARY

This report describes the technical progress accomplished under Air Force Office of Scientific Research (AFOSR) funding (grants F49620-01-1-0261and F49620-03-1-0024) spanning the performance period of February 2001 through December 2003. This effort was initiated at New Mexico State University (NMSU) under the first award and titled "Team Cognition in Distributed Mission Environments." It was continued at Arizona State University (ASU) East where the principal investigator relocated, along with the CERTT (Cognitive Engineering Research on Team Tasks) Lab in January 2003 under the second award and title, "The Role of Individual and Team Cognition in Uninhabited Air Vehicle Command-and-Control." This report documents the research conducted in the total 34-month effort.

The original goal of this project was to empirically examine the effects of distributed mission environments (vs. co-located environments) and workload on team performance, process, and cognition. In the first experiment we noted virtually no degradation of team performance in the distributed mission environment. Instead we noticed across team performance variance that could be attributed to the characteristics of the individuals on the team (e.g., gender, working memory capacity). We decided to better control for these differences in the second experiment and to carefully examine data collected in these and two previous experiments in order to identify individual characteristics and those associated with the various team roles that were predictive of team performance. In this archival analysis methodological issues are also addressed including reliability and validity of our measures of team cognition. In addition to the archival analyses, we collected data from expert teams in order to establish a performance benchmark.

These studies were conducted in the context of a UAV (Uninhabited Air Vehicle) ground control simulation in the CERTT Laboratory. The simulation focuses on the cognitive and team aspects of ground control and involves three team members (AVO - Air Vehicle Operator, PLO - Payload Operator, and DEMPC - Data Exploitation, Mission Planning, and Communications Operator). The team members interacted over headsets and computers in order to take reconnaissance photos of targets.

In the two dispersion (co-located vs. distributed) experiments, twenty teams (ten in each dispersion condition) participated in four low workload missions, followed by high workload missions (three in the first and one in the second experiment). Results generally indicated effects of workload on performance, process, and cognition (i.e., situation awareness). However there were no effects of dispersion on team performance, although this factor did affect cognition and team process. Distributed teams seemed to adapt to environments associated with less knowledge sharing by coordinating differently. Later analyses of communications (under a separate Office of Naval Research (ONR)-funded effort) support this claim.

A third small study in the CERTT Lab was undertaken with five expert teams in order to better understand the upper limits of performance, process, and cognition on this task. Teams were experienced at working together in technological or aviation environments. Results indicated that the highest scoring teams included a team of CERTT Lab experimenters who were capable of "gaming" the system and a team with extensive internet video game experience. The highest scoring expert teams also exhibited team process behaviors superior to lower scoring teams.

Results from the archival data analysis focusing on characteristics of individual team members or roles indicate that individual levels of situation awareness are positively correlated with level of team performance and that some factors such as situation awareness, working memory capacity, and grade point average are associated with performance for specific team positions (i.e., roles – AVO, PLO, DEMPC).

The evaluation of our new methods for measuring team cognition indicates that with the exception of teamwork knowledge, all of our primary measures have adequate reliability. Further, our situation awareness measure had significant predictive validity. In addition our holistic knowledge measures generated distinct results from our collective knowledge measures and both contributed to the variance accounted for in team performance.

Overall this work contributes to applications, theory, and methodology associated with team cognition. An important applied finding is that distributed mission environments seem to have minimal detrimental impact on team performance, at least for command-and-control tasks like the one tested here. Although knowledge and process are affected by geographic dispersion, teams seem to adapt to distributed environments by modifying the coordination process. These results provide encouraging support for the concept of network centric warfare. From a theoretical perspective, this work has shed light on the importance of this coordination process to team cognition. Team coordination plays a central role in team cognition. Much like cognitive processes operate on knowledge at the individual level, team coordination operates on knowledge at the team level. It appears then from these studies that team cognition is largely the team's capabilities for pushing and pulling knowledge in the form of information. Our current efforts in the CERTT Lab focus on understanding the development and retention of team coordination through empirical and modeling efforts. Finally, from a methodological perspective, this effort has led to the development and testing of appropriate methods for assessing knowledge at the team level (i.e., holistic methods). It has also revealed weaknesses in the knowledge measures that can be targeted for future improvements. Most importantly this project suggests that what lies at the heart of team cognition is not so much what is in the heads of the team members as it is in their interactions.

2.0 RESEARCH TEAM

Principle Investigator
Nancy J. Cooke (ASU)

Post Doctoral Assistant
Brian G. Bell

Graduate Students
Olena Connor (NMSU*)
Janie DeJoode (NMSU*)
Pat Fitzgerald (ASU)
Rebecca Keith (NMSU)
Harry Pedersen (NMSU*)

Undergraduates
Paulette Dutcher (ASU)

Subcontractor/CERTT Developer:
US Positioning: Steven M. Shope

Associated Personnel
Faculty
Peter Foltz (NMSU)
Graduate Students
Jamie Gorman (NMSU*)
Preston Kiekel (NMSU*)

*These students relocated from NMSU to ASU, though still officially working on NMSU degrees.

3.0 INTRODUCTION

3.1 The Problem

Technological developments in the military and elsewhere have transformed highly repetitive manual tasks, requiring practiced motor skills to tasks that demand cognitive skills often related to overseeing new technology such as monitoring, planning, decision making, and design (Howell & Cooke, 1989). As a result, a full understanding of many tasks, at a level required to intervene via training or system design, requires an examination of their cognitive underpinnings. Additionally, the growing complexity of tasks frequently surpasses the cognitive capabilities of individuals and thus, necessitates a team approach. For instance, teams play an increasingly critical role in complex military operations in which technological and information demands necessitate a multioperator environment (Salas, Cannon-Bowers, Church-Payne, & Smith-Jentsch, 1998).

Whereas the team approach is often seen as a solution to cognitively complex tasks, it also introduces an additional layer of cognitive requirements that are associated with the demands of working together effectively. Team members need to coordinate their activities with others who are working toward the same goal. Team tasks often call for the team to detect and recognize pertinent cues, make decisions, solve problems, remember relevant information, plan, acquire knowledge, and design solutions or products as an integrated unit. Therefore, an understanding of team cognition, or what some have called the new "social cognition" (Klimoski & Mohammed, 1994), is critical to understanding team performance and intervening to prevent errors or improve productivity and effectiveness.

The assessment and understanding of team cognition (i.e., team mental models, team situation awareness, team decision making) requires psychometrically sound measures of the constructs that comprise team cognition. However, measures and methods targeting team cognition are sparse and fail to address some of the more interesting aspects of team cognition (Cooke, Salas, Cannon-Bowers, & Stout, 2000). In addition, to be applicable to complex multioperator military contexts, such measures need to be developed and evaluated in a task environment that is conducive to scientific rigor, yet applicable to the operational settings in which the measures are to be extended. Thus, we have identified as a long-term research goal the development and evaluation of measures of team cognition in a military context. At the same time, as measures of team cognition are developed they can be used to better understand, train and design for superior team cognition.

3.2 Long-Range Objectives

The goal described above, involving the development and evaluation of measures of team cognition in a military context, can be decomposed into the following long-range objectives:

- Develop a military synthetic task environment that emphasizes team cognition.
- Identify needs and issues in the measurement of team cognition.
- Develop new methods suited to the measurement of team cognition.
- Evaluate newly developed measures.

- Apply measures to better understand team cognition.
- Apply measures to evaluate interventions relevant to team cognition.

Since 1997, when our research program was first funded by AFOSR, we have made significant progress toward these long-range objectives. This progress is described after the following section in which we provide theoretical and methodological background for our research program.

3.3 Background

Our research program is based on, and contributes to research and theory in several different areas. Recently, research and theory in the team arena has expanded upon work in industrial organization psychology on teams, work in social psychology and management on small groups, and work in human-computer interaction on groupware to spawn a new research area referred to as *team cognition*. The definition of team cognition and various issues in its measurement are covered in the first section (3.3.1). Our work also assumes that the context of a job or task is relevant and so embraces the synthetic task paradigm as a means to conduct controlled experiments in a realistic setting. The second section (3.3.2) describes synthetic task environments. Whereas team cognition and synthetic tasks are concepts that underlie our entire research program, there are also two bodies of literature that are relevant to the specific research questions addressed under this effort. The first deals with issues associated with teams in distributed mission environments (3.3.3) and the second deals with the relationship between individual characteristics and team performance (3.3.4). Thus, in the following sections we provide some background information associated with each of these topics.

3.3.1 The Measurement of Team Cognition

Salas, Dickinson, Converse, and Tannenbaum (1992) define *team* as "a distinguishable set of two or more people who interact dynamically, interdependently, and adaptively toward a common and valued goal/object/mission, who have each been assigned specific roles or functions to perform, and who have a limited life span of membership" (p. 4). Thus, teams, unlike some groups, have differentiated responsibilities and roles (Cannon-Bowers, Salas, & Converse, 1993). This division of labor is quite common in military settings and enables teams to tackle tasks too complex for any individual. Interestingly, this feature is also one that has been neglected in current measurement practices.

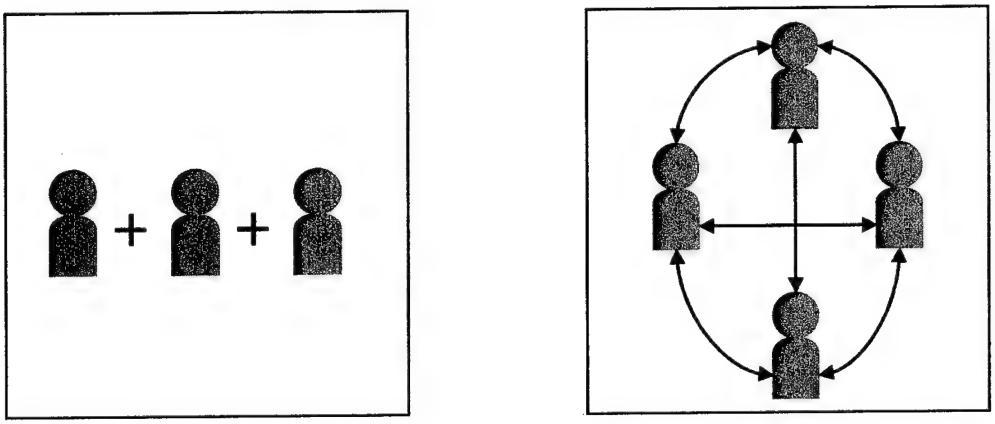
Team process behaviors such as communication, leadership behaviors, coordination, and planning have been linked theoretically and empirically to team performance (Foushee, 1984; Stout, Salas, & Carson, 1994; Zalesny, Salas, & Prince, 1995). Many interventions for improving team performance have targeted team process behavior (Braun, Bowers, Holmes, & Salas, 1993; Leedom & Simon, 1995; Prince, Chidester, Cannon-Bowers, & Bowers, 1992; Prince & Salas, 1993). Recently, it has become clear that other factors that are more cognitive than behavioral in nature also play a role in team performance. There has been significant theoretical work delineating cognitive constructs such as team decision making, shared mental models, and team situation awareness (Cannon-Bowers, et al., 1993; Orasanu, 1990; Stout, Cannon-Bowers, & Salas, 1996). It is assumed that with an understanding of these constructs, training and design interventions can target the cognitive underpinnings of team performance.

Also, the hypothesized relation between team cognition and team performance suggests that team performance in information-centric tasks can be predicted from an assessment of team cognition, thereby circumventing the need for teams to perform in less than optimal settings (e.g., minimal training, hazardous or high-risk environments) for performance assessment.

Our research on team cognition has, until recently, focused on team knowledge. Parallel to research on individual expertise (e.g., Chase & Simon, 1973; Glaser & Chi, 1988), accounts of effective team performance highlight the importance of knowledge, or in this case, team knowledge. For instance, Cannon-Bowers and Salas (1997) have recently proposed a framework that integrates many aspects of team cognition in the form of teamwork competencies. They categorize competencies required for effective teamwork in terms of knowledge, skills, and attitudes that are either specific or generic to the task and specific or generic to the team. Similarly, a team's understanding of a complex and dynamic situation at any one point in time (i.e., team situation awareness) is supposedly influenced by the knowledge that the team possesses (Cooke, Stout, & Salas, 1997; Stout, et al., 1996).

Based on this theoretical work and our own observations, we have developed a framework (see Figures 1a and 1b) that helps to better define team knowledge, and especially, to distinguish team knowledge as it has been traditionally measured (i.e., collectively) from team knowledge as it may best be measured (i.e., holistically). Traditional collective measurement (Figure 1a) involves eliciting knowledge from individuals on the team and then aggregating the individual results to generate a representation of the collective knowledge of a team. Although we believe that knowledge measured collectively should be predictive of team performance, it is also an oversimplification, devoid of the influences of team process behaviors (e.g., communication, coordination, situation awareness). These process behaviors are analogous to individual cognitive processes in that they transform the collection of team member knowledge into effective knowledge that is associated with actions and ultimately, with team performance in a dynamic environment. One of our research goals is to identify ways to measure effective team knowledge using team-level or holistic metrics (Figure 1b). Further, it is questionable whether simple aggregation of individual team member knowledge is appropriate for a team of individuals who have different roles and consequently, different knowledge bases. Although not depicted in Figure 1, team knowledge is multifaceted and consists of background knowledge that is long-lived in nature, as well as more dynamic and fleeting understanding that an operator has of a situation at any one point in time. Measures of team cognition have focused primarily on the former, at the expense of the latter.

Reliable and valid measurement of constructs like team knowledge is a first, albeit nontrivial step, that presents a "road block" to advances in our understanding of team cognition. Many parallels can be drawn between the measurement of individual and team cognition, given that the primary difference is whether the measurement is directed at the team or individual. Just as individual cognition is reflected in the behavior of the individual, team cognition is reflected in the behavior of the team. However, our focus on team knowledge measurement (most closely aligned with the shared mental model literature) has highlighted several areas in which measurement can be improved, including the tendency for researchers to target team cognition by focusing on the individual level and then aggregating results.



Panel A **Panel B**

Figure 1. Panel A represents collective approaches to the measurement of team knowledge. Panel B represents holistic approaches to the measurement of team knowledge which consider team process behaviors as integrators of individual cognition.

Historically, measures of team cognition tend to explore a small portion of the space of possible measures as is indicated in Table 1. This table classifies team knowledge measures according to type (long-term, fleeting) and the metric used to assess the knowledge elicited from individuals. For the most part, researchers have looked at intrateam similarity of knowledge structures and accuracy of those structures with regard to some referent. There are other possible classification schemes not included here such as whether the knowledge is declarative, procedural or strategic, the type of technique used to elicit the knowledge in the first place, and whether the elicitation is collective or holistic. The Xs in the table indicate the cells in which measurement work has taken place with large Xs indicating more work. Apparently there remains much room for further development.

Table 1
Current Progress in Measures of Team Knowledge

TYPE OF KNOWLEDGE	METRIC			
	Similarity	Accuracy	Role Accuracy	Interpositional Knowledge
Long-term (shared mental models)	X	X		
Dynamic (team situation models)	x	x		

The various measurement issues relevant to team knowledge that have been identified thus far are described in detail in Cooke, et al., (2000) and are briefly summarized in Table 2 below.

Table 2
Issues in the Measurement of Team Cognition

• Measures are needed that target the holistic level, rather than the collective level, of team cognition (i.e., elicit team knowledge from the team).
• Measures of team cognition are needed that are suited to teams with different roles (e.g., navigator, pilot).
• Methods for aggregating individual data to generate collective knowledge need to be investigated.
• Measures of team knowledge that target the more dynamic and fleeting situation models are needed.
• Measures that target different types of team knowledge (e.g., strategic, declarative, procedural knowledge or task vs. team knowledge) are needed.
• The extension of a broader range of knowledge elicitation methods to the problem of eliciting team cognition is needed.
• The streamlining of measurement methods to better automate them and embed them within the task context is needed.
• The validation of newly developed measures is needed.

3.3.2 Synthetic Task Environments

Our work has been greatly influenced by the assumption that synthetic tasks provide ideal environments for cognitive engineering research on complex tasks. We have developed an STE (Synthetic Task Environment) based on the real task of controlling a UAV. Our research and methodological developments in team cognition take place in this context.

Synthetic tasks are "research tasks constructed by systematic abstraction from a corresponding real-world task" (Martin, Lyon, & Schreiber, 1998, p. 123). Performance on a synthetic task should exercise some of the same behavioral and cognitive skills associated with the real-world task. An STE provides the context for a suite of synthetic tasks. This environment offers a research platform that bridges the gap between controlled studies using artificial laboratory tasks and uncontrolled field studies on real tasks or using high-fidelity simulators.

An STE can be considered a type of simulation, but philosophically differs from traditional simulations in terms of goals and resulting design decisions. Simulations typically recreate the work environment and the equipment or systems within that environment. An STE is "task centric" in that the goal is to recreate aspects of the task to differing degrees of fidelity. Thus, an STE may not have the "look and feel" of the operational environment, but instead calls upon the same cognitive structures and processes of the operational task. Because tasks are often situated in rich environments, STEs often include simulations of systems required to support the task. However, the focus is on abstracting task features consistent with the purpose of the planned research associated with the STE and concomitant design objectives. Thus, verisimilitude is maximized for aspects of the task under direct study. As a result, several very different STEs can be based on the same real task by virtue of applying distinct filters, each associated with different objectives. Such is the case with the UAV task in which a variety of STEs have been

developed that focus on various cognitive skills of individuals (e.g., Gugerty, Hall, & Tirre, 1998; Martin et al., 1998) and others, such as our UAV-STE, focusing on team cognition. In addition, simulations often replicate the environment at the expense of the simulation's flexibility as a research tool. Researchers are limited in the degree to which they can alter or control the simulation and the measures that they can derive from it. STEs, on the other hand, typically incorporate multiple task scenarios, and often the ability to manipulate aspects of task scenarios, as well as flexibility in measurement. This increased flexibility is not so much inherent in the concept of an STE, as demanded by researchers who appreciate the benefit of trading off some aspects of fidelity for research flexibility (e.g. Fowlkes, Dwyer, Oser, & Salas, 1998; Cannon-Bowers, Burns, Salas, & Pruitt, 1998). Recently, researchers have cautioned against the use of simulations unguided by training principles or an understanding of the actual task requirements and have extolled the virtue of low-fidelity simulations that take such factors into account (Miller, Lehman, & Koedinger, 1999; Salas, Bowers, et al., 1998).

Synthetic task environments, like high-fidelity simulations, can facilitate research in a safe and inexpensive setting and can also be used for task training and system design in support of tasks. They are also touted as providing a viable middle ground between overly artificial lab research and uncontrollable field research (Brehmer & Dorner, 1993). In many ways, STEs provide the best of both worlds -- the laboratory and the field. Alternatively, if they fail to meet the combined objectives of experimental control and sufficient representation of the task in question, they may instead capture the worst of both worlds—poor experimental control and low fidelity.

Whereas lack of experimental control has not been a major criticism levied against STEs, lack of fidelity has. STEs have been described as *low-fidelity* simulations, as opposed to traditional equipment-centric simulations. Indeed, STEs may have low fidelity in terms of replicating the features of the equipment. The low fidelity criticism is tied to more general concerns about low face validity. This issue is addressed by Salas, Bowers, et al. (1998), however, who argue that face validity may dictate acceptance by users, but not necessarily success as a training or research tool.

Perhaps more importantly, low fidelity is linked to low external validity and consequently, lack of generalizeability to the situation of interest. On the other hand, this low external validity criticism breaks down if fidelity is considered more broadly. Fidelity is generally the match between the research environment and the specific environment to which results are assumed to transfer. The match, however, can be based on a number of dimensions including the equipment and the task requirements. Thus, fidelity is not necessarily a single feature that is high-or-low for a particular simulation, but rather a multidimensional feature that can ultimately result in contexts of mixed fidelity. That is, a simulation may be faithful to the equipment, but not to the task requirements. In light of the issue of external validity, it is important to determine the dimensions of the transfer situation that are relevant to the research questions to be generalized. A mixed fidelity simulation may have adequate external validity, and thus generalizeability to the actual setting, if it is faithful to the relevant dimensions of the actual setting. Determining external validity then becomes a question of accurately identifying the relevant dimensions in the field of practice for the research questions. Generalizing results to other settings amounts to identifying similar features along the same relevant dimensions in those settings. It can then be assumed that the match is sufficient for research results to generalize to this environment. This

enterprise of identifying and matching the features and dimensions among different work environments amounts to a theory of tasks or transfer across work environments.

Under this multidimensional view of fidelity, the labeling of traditional simulations as *high fidelity*, and of STEs as *low-fidelity*, makes little sense. Instead, STEs are typically high fidelity with respect to the task and low-fidelity with respect to the equipment. Traditional simulations may more often be associated with the opposite pattern. External validity cannot be determined independent of the research questions. Research on cognitive aspects of a complex task such as decision making under stress, may best be addressed in a context that preserves the features of the task at the expense of the fidelity of the equipment. Alternatively, research directed at uncovering reasons for operational errors associated with a piece of equipment may generalize only in a context that faithfully replicates the equipment, perhaps at the expense of simplifying the overall task. The question of the external validity and extent of generalizeability of both traditional simulations and STEs needs to be addressed for each test-bed in the context of each research question.

3.3.3 The Problem of Team Cognition in Distributed Environments

In the course of conducting research in the UAV-STE, we observed in our co-located teams a need to discuss the specific jobs of the other team members during break periods, as well as a desire to view the computer displays of fellow team members. Overall, the co-presence of team members between missions seemed to be an important factor for this task.

In contrast, today's military tasks are performed by teams of individuals who may have never met each other; who are not necessarily sitting together in the same briefing room or side-by-side on the same battlefield; and who may only communicate and share information via communication and computer technologies. Indeed, the entire nature of warfare has taken on a "network centric" characteristic (Wilson, 2000) in which the battlefield is dispersed not only over terrain, but also over the internet. Further, this holds not only for teams of individuals, but also for teams of teams or hierarchical teams in which the task is shared by an intricate *distributed* network of collaborating individuals who share large amounts of information.

Before further describing the nature of the problem, it is important to clarify our use of the term "distributed." Recently, some investigators have characterized team cognition as "distributed" cognition, in the sense that cognition is dispersed over an entire sociocultural system (Hutchins, 1991; Rogers & Ellis, 1994). Furthermore, team expertise has been characterized as "distributed" in that team members each have different backgrounds and skills to contribute to the team goals (Hollenbeck, Ilgen, Sego, Hedlund, Major, & Phillips, 1995). Information may also be distributed in terms of time with communication occurring asynchronously as opposed to synchronously. Although team cognition and expertise may very well be distributed across team members, even in the same physical location, our use of the term "distributed" applies specifically to the physical location of the team members themselves. Team members in distributed environments are geographically dispersed.

How does the DME (Distributed Mission Environment) affect task performance by individuals, teams, and teams of teams? In comparison to contexts in which team members are co-located, in DMEs team members are less likely to be familiar with one another, must often communicate in

ways other than face-to-face communication, and due to lack of co-presence may not easily share displays or information conveyed through gestures. Such factors are likely to affect team process behaviors such as communication, coordination, and planning, which in turn, should affect team performance, even for skilled teams under conditions of high workload (Robertson & Endsley, 1997). In addition, we speculate that the process limitations associated with DMEs also affect team cognition (i.e., team decision making, team situation models, and shared mental models) by virtue of the relation between process and team knowledge and situation awareness depicted in Figure 1. For instance, communication breakdowns can affect the ability of a team to develop a shared understanding of the task and of the immediate situation. Given the intense information sharing and communication requirements of typical DME tasks, it is likely that team cognition plays a significant role in team performance in such environments. If dispersion affects team cognition and assuming team cognition is a vital contributor to team performance, then team performance should benefit from interventions at the cognitive level. In summary, we generally predict that process factors associated with DMEs will lead to deficits in the development of team cognition (and consequently team performance during acquisition) and later deficits in skilled performance.

Recognizing the differences between DMEs and traditional co-located team environments, DMT (Distributed Mission Training) has become of central importance to the Air Force (Wilson, 2000). Plans for DMT include distributed interconnected military simulations that present opportunities for extensive training in the DME. Interestingly, unlike simulations of physical battlefield maneuvers, simulations associated with DMT offer a high degree of face validity and realism to the training. Unfortunately, the scientific research base addressing pertinent issues of team cognition, its acquisition, and its assessment in this environment is notably sparse. Although Kleinman and Serfaty (1989) studied a military-based synthetic task environment that represented a geographically distributed AWACS (Airborne Warning and Control System) environment, there was no direct comparison of team behavior, performance, and cognition in this setting with a co-located environment. Generally, little is known about the effects of DMEs on team performance, behavior, and cognition.

Although there has been little or no research on the impact of DMEs on team performance or cognition in military settings, research in the human-computer interaction community has addressed mode of communication, a topic relevant to DMEs. This research has compared face-to-face or audio communication with computer-mediated communication, such as e-mail, GDSS (Group Decision Support Systems), or other tools. Several of these studies have found problems with computer-mediated communication. For example, Mantovani (1996) found that computer-mediated communication can hinder the creation of meaning and Hedlund, Ilgen, and Hollenbeck (1998) found that computer-mediated communication can lead to lengthy decision-making when compared to face-to-face communication. Unfortunately, these studies did not measure the performance of heterogeneous teams working on complex tasks for an extended period of time during which team member familiarity may confer an advantage. Therefore, these studies are limited in what they can tell us about the effects of DMEs on team cognition and performance because military DMEs are influenced by factors other than communication mode, such as team member familiarity and co-presence. Furthermore, military teams are composed of individuals with different roles as opposed to homogenous groups of individuals making the same judgment or decision.

Some researchers have found that specific group norms determine the impact of a DME, with higher team member familiarity among co-located teams producing better performance in comparison to DME teams (Postmes & Spears, 1998; Postmes, Spears & Lea, 1998; Contractor, Seibold, & Heller, 1996). Therefore, in two of the experiments that are reported here we compared co-located and distributed teams on several measures of team performance and team cognition. In addition to possible differences in team member familiarity affected by the distributed or co-located environment, potential differences in team member co-presence (including the ability to view the work environments of other team members) and communication mode (no face-to-face interaction for distributed teams) were expected to influence team performance, process, and cognition.

3.3.4 Individual and Team Cognition

Not only do teams detect and recognize cues, solve problems, and perform other cognitive functions as an integrated unit, but they also rely on the skills, knowledge, and abilities of the individuals who compose the team whenever these functions are performed. Kyllonen (1996) presents a general framework for understanding individual differences in cognitive ability that identifies various cognitive variables that differ at an individual level. The Cognitive Abilities Measurement (CAM) framework lists four major components of the human information-processing system: working memory, declarative knowledge, procedural knowledge, and processing speed. Working memory is a short-term store with limited capacity and contains what is currently available for processing at a given point in time. Processing speed refers to the time it takes to transform incoming information into motor responses. According to Kyllonen, each of these components can be divided into three domain areas: verbal, quantitative, and spatial. Kyllonen (1996) has found that working memory capacity, as a general factor composed of verbal, spatial, and quantitative components, accounts for much of the variance in learning various skills such as computer programming.

A few researchers (Barnes, Knapp, Tillman, Walters, & Velicki, 2000; Dolgin, Kay, Wasel, Lamgelier, & Hoffmann, 2001; Fatolitis, 2003) have recently examined the relationships between various cognitive and psychomotor ability measures and the performance of pilots as well as UAV operators. In their review, Dolgin et al., (2001) reported that in one study, between 30% and 45% of the variance in the number of flight violations by pilots could be accounted for by measures that assessed working memory capacity, mental flexibility, and divided attention. Fatolitis (2003) found significant correlations between various psychomotor and cognitive ability variables and the training performance of UAV operators, with a measure of visiospatial working memory capacity having the highest correlation among the measures. Barnes, et al. (2000) measured cognitive abilities of UAV operators using JASS (job assessment system software) and found that the requisite abilities suggested that operators need not be rated aviators. This work also revealed the importance of cognitive skills in the aviation domain as mentioned previously. However, these studies did not examine the performance of the entire team or did so by relying on very high-level outcome measures such as frequency of accidents and incidents.

What is relatively unknown is the impact of such individual differences on team performance and cognition. An early review (Heslin, 1964) examined whether the cognitive ability of team members was related to team performance. Heslin found in most of the studies that there was a positive correlation between general cognitive ability, as assessed by college grades or test

scores, and team performance. More recent studies, (LePine, Hollenbeck, Ilgen, and Hedlund, 1997; Hollenbeck, Moon, Ellis, West, Ilgen, Sheppard, Porter, and Wagner, 2002) have also found that greater cognitive ability is associated with better team performance.

The examination of heterogeneous teams in which members have different jobs, backgrounds, or roles, raised the interesting question about the relationship between individual characteristics tied to particular roles and team performance. For example, preliminary results in our lab have shown in the context of the UAV-STE that DEMPCs with more working memory capacity perform better, and that the performance of the DEMPC further has a strong influence on team performance. In this case, the role of the team member is also a factor in the relationship between individual abilities and team performance. By considering such differences among individuals, team roles, and between teams composed of different individuals (i.e., team differences due to team composition), we should be able to account for variance in team and individual performance that would be otherwise unexplained. By identifying and removing this variance, this approach may allow us to detect more subtle effects of manipulations on team performance. Additionally, identifying individual differences that are critical for team performance is a necessary step toward improving team performance through team composition, focused training, and design aids.

Focusing on individual and role-related cognitive skills and abilities as a means of understanding team cognition again raises a number of issues concerning measurement and aggregation. Although these variables can be analyzed at the individual level, they can also be combined to produce a team score (e.g., team working memory), which requires decisions about the best approach for deriving such a score. For example, is it more appropriate to average the individual scores to arrive at a team score or is another approach, such as taking the highest-performing member's score, a better estimate? Perhaps some variables are important for performance in some team roles, but not in others. As previously discussed, issues of heterogeneity, aggregation, and holistic measurement are all relevant here.

3.3.5 Background Summary

Our research program in general and the research reported here integrates research and theory from several different areas. Our goal to improve the measurement of team cognition draws from various literatures on teams, small groups, and individual cognition. Our use of the STE paradigm is driven by some philosophical considerations concerning basic vs. applied research and the limitations of each. Finally, the specific work presented in this report draws from problems observed in operational environments, some sparse literature on group dispersion, and cognitive individual differences.

3.4 Prior Progress Toward Long-Range Objectives

Our research program on team cognition was initiated in 1997 with a DURIP (Defense University Research Instrumentation Program; F49620-97-1-0149) grant that provided funds for initial equipment in the CERTT (Cognitive Engineering Research on Team Tasks) Laboratory. Subsequent grants from AFOSR (F49620-98-1-0287; F49620-01-1-0261, F49620-03-1-0024) have funded research in the CERTT Lab from 1998 to the present (2003). Our progress toward

the long-range objectives of our research program fall into three major areas: 1) CERTT Lab and UAV Synthetic Task Development, 2) Methodological Developments, and 3) Empirical Findings. This progress is summarized in the sections that follow.

3.4.1 CERTT Lab and UAV Synthetic Task Development

The CERTT Lab is a research facility for studying team performance and cognition in complex settings and it houses experimenter-friendly equipment to simulate these settings. Our work has been greatly influenced by the assumption that synthetic tasks provide ideal environments for cognitive engineering research on complex tasks in that they serve as a middle ground between the difficult to control field and the artificial tasks typically found in the lab. We have developed in the CERTT Lab a UAV-STE based on a cognitive task analysis (Gugerty, DeBoom, Walker, & Burns, 1999) of ground control operations for the Predator at Indian Springs, NV (Cooke, Rivera, Shope & Caukwell, 1999; Cooke & Shope, in press; Cooke & Shope, 2002a; Cooke & Shope, 2002b; Cooke & Shope, 1998; Cooke, Shope, & Rivera, 2000). This UAV-STE emphasizes team aspects of the task such as planning, replanning, decision-making, and coordination. Thus, our research and methodological developments in team cognition have taken place in this context. We believe that our research and methods relevant to team cognition in this environment can be generalized to other command-and-control environments.

CERTT's UAV-STE simulates a three-team member task in which each team member is provided with distinct, though overlapping training; has a unique, yet interdependent role; and is presented with different and overlapping information during missions. The overall goal is to fly the UAV to designated target areas and to take acceptable photos at these areas. The AVO controls airspeed, heading, and altitude, and monitors UAV systems. The PLO adjusts camera settings, takes photos, and monitors the camera equipment. The DEMPC oversees the mission and determines flight paths under various constraints. To successfully complete a mission, the team members need to share information with one another in a coordinated fashion. Most communication is done via microphones and headsets, although some involves computer messaging. Measures taken include audio records, video records, digital information flow data, embedded performance measures, team process behavior measures, situation awareness measures, and a variety of individual and team knowledge measures. The participant and experimenter consoles are depicted in Figures 2 and 3, respectively. Team members require 1.5 hours of PowerPoint and hands-on training before they are ready to interact as a team.

Figure 2. CERTT Lab participant consoles.

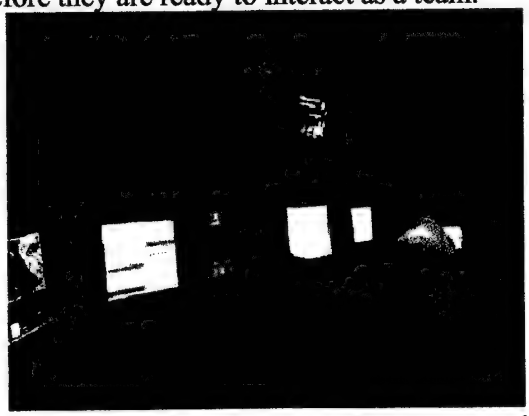


Figure 3. CERTT Lab experimenter console.

Features of the CERTT UAV-STE include:

- Four participant consoles (including remote console)
- One experimenter workstation
- Integration of seven task applications over local area net
- Video and audio recording equipment (including digital audio)
- David Clark headsets for participants and experimenter
- Intercom and software for logging communications flow
- Embedded performance measures
- Computer event logging capabilities
- Experimenter ability to disable or insert noise in channels of communication intercom
- Experimenter access to participant screens
- Experimenter control capability of participant applications
- Easy to change start-up parameters and waypoint library that define a scenario
- Software to facilitate measurement of team process behaviors
- Software to facilitate situation awareness measurement
- Training software modules with tests
- Software modules for off-line knowledge measurement (taskwork ratings)
- Software for administering debriefing questionnaire
- Software for administering NASA TLX (National Aeronautics and Space Administration Task Load Index), SART (Situation Awareness Rating Technique), and other scales
- Capability for distributed simulation (across intranet and internet)

3.4.2 Methodological Developments

Given that we have a long-term goal of developing and evaluating measures of team cognition and performance, many of our accomplishments are methodological in nature. Our methodological work and the various measurement issues relevant to team knowledge that have been identified thus far are described in detail elsewhere (Cooke, Kiekel, Bell, & Salas, 2002; Cooke, Kiekel, & Helm, 2001; Cooke, Shope, & Kiekel, 2001; Cooke, Stout, & Salas, 2001; Cooke, et al., 2000; Cooke, Stout, Rivera, & Salas, 1998; Cooke, et al., 1997).
Our methodological progress has included the development of training and measurement modules that interface with the CERTT Lab's UAV-STE including:

- UAV-STE waypoint database to facilitate scenario changes
- Communication flow logging software
- Participant performance score viewer and experimenter interface
- Software measures of working memory capacity and social desirability
- Critical incident and summary measures of team process behavior
- Systems for randomizing and recording responses to embedded situation awareness probes

We have also made methodological progress in developing and evaluating metrics that are more appropriate for the heterogeneous command-and-control teams that we study. We have developed:

- Holistic or consensus-based methods of measuring taskwork knowledge, teamwork knowledge, and situation awareness at the team level
- Accuracy metrics for heterogeneous teams that can quantify overall, positional, and interpositional accuracy of knowledge
- Proportion of agreement metrics
- Various aggregation schemes more appropriate for command-and-control than averaging responses
- Communication analysis as an unmitigated approach to the measurement of team cognition (funded by ONR, N00014-00-1-0818 and N00014-03-1-0580)

3.4.3 Empirical Findings

Prior to the current effort, two experiments (AF1 and AF2) were run using the UAV-STE under previous AFOSR effort (1998-2000). In addition, a student's M.A. thesis on collaborative writing was also conducted using the CERTT equipment, as was an M.S. master's project on individual data aggregation. The first UAV-STE experiment (AF1) examined acquisition of team performance in this environment with eleven 3-person teams performing ten 40-minute missions. The second experiment (AF2) compared ten teams in environments conducive to knowledge sharing to eight teams in an environment not conducive to such sharing (i.e., no talking about the task allowed or looking at others' computer displays). These experiments were conducted to evaluate and iterate on the UAV-STE, to test newly developed measures of team cognition, and to begin to understand some of the factors relevant to team cognition.

This early empirical work in the UAV-STE context and in other STE contexts (a Navy low-fidelity helicopter simulation) not only aided iterative design of our task, experimental procedures, measures, and training materials, but also generated the following findings regarding team cognition and associated measures:

- Team performance reaches asymptotic levels after four 40- minute missions.
- Several knowledge measures/metrics are predictive of team performance in this context.
- In this context, interpositional accuracy tends to develop with task and team experience (i.e., good team members are not specialists).
- Taskwork knowledge is relatively stable after initial task training, whereas teamwork knowledge tends to develop with mission experience.
- Early attempts to force-feed teamwork or coordination information prior to development of taskwork knowledge have not succeeded, suggesting a sequential dependency in knowledge development (taskwork must precede teamwork).
- Encouraging or discouraging information sharing during breaks and by examining others' displays had no effect on team performance.

Of particular interest, in the UAV-STE individual team members are able to quickly (1.5 hours) acquire the skill that they needed to perform their individual roles. Team performance, however, as measured by a composite score made up of components relevant to the rate of performance (e.g., number of targets successfully photographed per minute) develops to asymptotic levels over four 40–minute missions after individual training (see Figure 4). Team situation awareness followed a parallel developmental path. This pattern of skill acquisition on this team task has

been replicated across the experiments that we have conducted. Because individuals have attained a criterion level of performance prior to the first mission as a team, it is team skill that develops over the first four missions. In particular, we assume that team members are learning how to coordinate or share information with the right person at the right time.

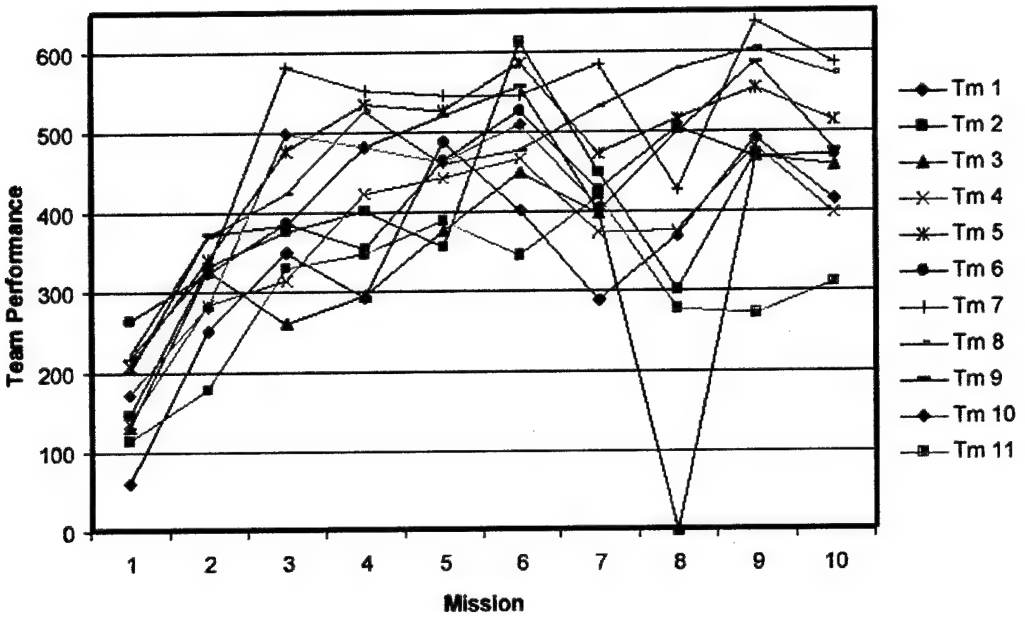


Figure 4. Acquisition of UAV task (team performance scores) for 11 teams in Experiment AF1.

Further, the teams' skills were not specific to the particular UAV scenario (i.e., novel scenarios were introduced for Missions 7 and 10) in so much as performance was unaffected by a novel scenario. Performance was, however, affected by a long break of several weeks, especially for lower performing teams. Finally, although the best teams in Experiment AF1 had knowledge that resembled a global view of the task (i.e., from the perspective of all three team positions), attempts to directly train individuals so that this form of knowledge would be acquired (in Experiment AF2, shared knowledge condition) succeeded in terms of team knowledge acquisition, but had no impact on performance. Thus, it seems that the possession of a global view of the task is only partly responsible for high levels of team performance. It is likely that both global knowledge and team process behaviors play a role in team performance and that mastery of the process component of skill was thwarted by the knowledge manipulation in AF2. In other words, teams who were force-fed taskwork knowledge may have missed out on the development of adequate process skills that allowed the low-knowledge teams to compensate. In general, the UAV-STE provides a complex and dynamic task environment in which teams can reach proficiency in a reasonable amount of time, yet teams can also be differentiated from each other in terms of their level of skill and concomitant knowledge and process.

These two empirical studies have served to identify promising methods for measuring team cognition. In general, the taskwork relatedness rating measures taken at the individual level seem to provide useful information about the team's knowledge of the task from the perspective

of each team role. The teamwork questionnaires used in these experiments reflected knowledge that changed with mission experience, but was not generally associated with team performance. The measure of team situation awareness, on the other hand, seemed to capture the momentary knowledge of the team regarding the mission in progress. This measure was predictive of performance in both experiments and unlike the taskwork rating-based measures, was administered in the form of experimenter queries randomly interspersed through the mission. Other methods tested in these experiments have not been as successful at measuring that which they were intended to measure, including the taskwork consensus ratings, the taskwork questionnaire, and the team process measure.

Further these empirical studies have led to some refinements of our framework for understanding team cognition. For instance, though we viewed team process behaviors as central to team cognition in our earlier conception, we now view team process behaviors as cognitive processing at the team level. The fortunate aspect of this new conceptualization is that unlike individuals, we can directly observe cognitive processing at the team level, opening the door to numerous measurement possibilities.

3.4.4 Summary of Early Contributions

In summary, progress prior to this effort took steps toward achieving the long-range objectives that we have established. Specifically we made several contributions in this early work:

- The development of a facility (i.e., CERTT Lab) dedicated to cognitive engineering research on team tasks
- The development of an STE for teams, based on UAV operations
- The design of an interface for UAV operations that requires much less training time than the Predator ground control interface
- Explicit procedures for designing and developing an STE
- Identification of issues and problems related to measuring team cognition
- New measures of team knowledge that look promising in terms of their predictive validity
- New metrics to aid in measuring team cognition in heterogeneous teams
- Empirical studies investigating team cognition and the acquisition of team skill on the UAV task
- A suggestion that cross training may work because of its focus on taskwork training before teamwork training
- A conceptual framework for understanding team cognition

3.5 Objectives of Current Effort (2001-2003)

The objectives and tasks listed below combine those objectives and tasks across the two associated efforts (F49620-01-1-0261and F49620-03-1-0024) spanning the 34-month period from 2001 to 2003. The objectives include empirical studies to address effects of geographic dispersion and workload, empirical efforts to benchmark performance in the CERTT UAV-STE, and archival data analyses to investigate the role of individual characteristics on team performance and to further evaluate measures of team cognition.

These seemingly disparate objectives evolved over the course of the three-year project. The original goal of this project was to empirically examine the effects of distributed mission environments (vs. co-located environments) and increased workload on team performance, process, and cognition. However, in the first experiment we noted virtually no degradation of team performance in the distributed mission environment. Instead we noticed across team variance that seemed to be attributed to the characteristics of the individuals on the team (e.g., gender, working memory capacity). We decided to better control for these differences in the second experiment and to carefully examine data collected in these and two previous experiments to identify individual characteristics and those associated with team roles that are predictive of team performance. In this archival analysis methodological issues are also addressed including reliability and validity of our measures of team cognition. In addition we collected data from expert teams in order to establish a performance benchmark.

Objective 1: Empirical Studies on DMEs. Conduct empirical studies to investigate the impact of geographic dispersion and varying workload on team performance, process, and cognition in the context of the CERTT Lab's three-person UAV-STE.

- **Task 1:** Design and collect data from an experiment to investigate the combined effects of communication mode differences, familiarity, and co-presence on team cognition, process, and performance during task acquisition and skilled performance under varying workload conditions.
- **Task 2:** Analyze data from the first experiment to determine the direction for the follow-up experiment.
- **Task 3:** Based on the data from the first study, design and collect data from a second experiment to investigate the combined effects of communication mode differences, familiarity, and co-presence on team cognition, process, and performance during task acquisition and skilled performance under varying workload conditions.

Objective 2: Empirical Study to Benchmark Performance. Conduct an empirical study to benchmark expert performance in the context of the CERTT Lab's three-person UAV-STE.

- **Task 1:** Determine requirements for expert teams and recruit expert teams for experiment
- **Task 2:** In a single session with five missions collect performance, cognitive, and process data from expert teams
- **Task 3:** Compare data from expert teams to previously collected data from non-expert teams.

Objective 3: Archival Analysis of Individual and Role-Associated Characteristics. Investigate the relation between individual characteristics and team cognition and performance through an archival analysis on data from four previously conducted CERTT UAV-STE experiments.

- **Task 1:** Assemble data collected from four CERTT-UAV experiments
- **Task 2:** Across four experiments, attempt to identify individual and team differences (cognitive and otherwise) that account for significant variance in team performance.

- **Task 3:** Explore the impact of individual differences associated with team role on team performance and cognition
- **Task 4:** Explore the use of voice stress as an index of individual arousal during mission performance

Objective 4: Archival Analysis to Evaluate Measures. Evaluate the newly developed measures and metrics of team cognition in terms of reliability and validity through an archival analysis on data from four previously conducted CERTT UAV-STE experiments.

- **Task 1:** Assemble data collected from four CERTT-UAV experiments
- **Task 2:** Evaluate across the four experiments measures of team cognition especially in terms of measure reliability and validity.
- **Task 3:** Conduct a multitrait multimethods (MTMM) analysis on data collected from the third experiment.
- **Task 4:** Examine the benefit of holistic vs. collective measures of team cognition across the four experiments.
- **Task 5:** Address aggregation of individual data for measures of team cognition at the collective level.

3.6 Our Approach

In each of the sections that follow we report for each objective our specific approach to the problem, hypotheses, methods, results, and conclusions. In general our approach to the first two objectives is an empirical one in the context of the UAV-STE. When manipulating dispersion in these studies, we do so while maintaining some fidelity in regard to typical communication mode. That is, UAV operators typically communicate over headsets while looking at computer screens, even when they are co-located. Thus, our co-located condition is not a true "face-to-face" condition in which communication is unimpeded by technology. We take this approach, however, to address more realistic variations of geographic dispersion.

Our benchmarking study is also approached through empirical data collection. Ideally we would have preferred to use intact teams of Predator UAV operators. If the UAV-STE *is* indeed faithful to the cognitive and team aspects of the operational task, then these operational teams would provide the best performance benchmark. However, subject matter experts are difficult to obtain, even in times of peace. In 2003 this problem was exacerbated due to the war in Iraq. We therefore decided to collect data from intact teams who had experience interacting in similar ways (i.e., ideally over headsets and through computers in a command-and-control like task).

Finally, questions about methodological validity and individual differences are difficult to address in a single small experiment with 20 or so teams. Therefore, our archival studies were designed to take advantage of the data collected over the course of four experiments. Together we have collected data on 69 teams in the CERTT UAV-STE context and these archival analyses are an attempt to find patterns in the data that are consistent across experiments.

4.0 PROGRESS UNDER THIS EFFORT

4.1 Experiment 1: Team Cognition in Distributed Mission Environments

Experiment 1 was designed to address the first objective of this project which was to *conduct empirical studies to investigate the impact of geographic dispersion and varying workload on team performance, process, and cognition in the context of the CERTT Lab's three-person UAV-STE.* The first task under this objective was to *design and collect data from an experiment to investigate the combined effects of communication mode differences, familiarity, and co-presence on team cognition, process, and performance during task acquisition and skilled performance under varying workload conditions.*

As described in the background section, the rationale for this experiment was based on observations of important co-located interactions during previous CERTT experiments, the importance of distributed environments in current military operations, interesting theoretical questions about distributed team cognition, and a dearth of empirical data on team performance in DMEs.

Given the intense information sharing and communication requirements of typical DME tasks, it is likely that team cognition plays a significant role in team performance in such environments. We generally predict that process deficits associated with DMEs will lead to concomitant deficits in the development of team cognition. Thus, if dispersion negatively affects team cognition and if team cognition is a vital contributor to team performance, then team performance should suffer in DMEs compared to co-located environments.

There will be two phases of the task: (1) task acquisition lasting for four missions; and (2) skilled task performance in which workload is increased for the last three missions. Measures of performance, process, and cognition will be taken as stated previously.

The following hypotheses are based on the assumptions stated previously regarding factors associated with DMEs, as well as our theoretical views concerning the relations between team cognition, process, and performance.

H1.1 During task acquisition DME teams will suffer process deficits resulting in slower acquisition rates and overall poorer acquisition performance compared to teams in the co-located condition.

H1.2 During task acquisition DME teams will suffer process deficits resulting in slower development of team knowledge and situation models compared to teams in the co-located condition.

H1.3 Although by later trials, DME teams may "catch up" in terms of team cognition and performance to co-located teams, and may compensate for process deficits during low workload periods, process deficits, and consequently performance and situation model deficits, will occur in periods of high workload.

H1.4 Individual differences among DME teams in terms of process strategy may moderate any deleterious effects of the DME, such that the "best" DME teams can overcome DME limitations compared to teams with poorer team process.

4.2 Experiment 1: Method

4.2.1 Participants

Twenty three-person teams of NMSU students voluntarily participated in two (5 hour) sessions for this study. Individuals were compensated for their participation by payment of $6.00 per person hour to their organization. The number of participants from each organization can be found in Appendix A. In addition, the three team-members on the team with the highest performance score were each awarded a $50.00 bonus.

Most of the participants were either Caucasian (55%) or Hispanic (27%) with males representing 65% of the sample. Participants ranged in age from 18 to 40. The participants were randomly assigned to teams and to role (AVO, PLO, or DEMPC). One team was replaced because a member of the team did not understand English. Teams were randomly assigned to either the co-located or distributed condition.

4.2.2 Equipment and Materials

The study took place in the CERTT Lab configured for the UAV-STE described previously. For most of the study, each participant was seated at a workstation consisting of two computer monitors (one View Sonic monitor connected to an IBM PC 300PL and one Dell Trinitron monitor connected to a Dell Precision 220 PC), a Sony video monitor that could present video from a Quasar VCR, two keyboards, and a mouse for input. Participants communicated with each other and the experimenters using David Clark headsets and a custom-built intercom system designed to log speaker identity and time information. The intercom enabled participants to select one or more listeners by pressing push-to-talk buttons.

Two experimenters were seated in a separate adjoining room at an experimenter control station consisting of four Dell Precision 220 PCs and Dell Trinitron monitors, an IBM PC computer and Panasonic monitor, two Panasonic monitors for viewing video output, and two Sony monitors for video feed from ceiling mounted Toshiba cameras located behind each participant. In addition, a fourth camera captured information from the entire participant room.

From the experimenter workstation, the experimenters could start and stop the mission, query participants together or individually, monitor the mission-relevant displays, select any of the computer screens to monitor using a Hall Research Technologies keyboard video mouse (KVM) matrix switch, observe team behavior through camera and audio input, and enter time-stamped observations. A Javelin Systems Quad Splitter allowed for video input from each of the 4 cameras to be displayed simultaneously on the monitor and was recorded on another Quasar VCR. In addition, a video overlay unit was used to superimpose team number, date, and real-time mission information on the video. Audio data from the headsets was recorded on an Alesis digital recorder and sent to the VCR and a Denon Precision Audio Component cassette recorder. Furthermore, custom software recorded communication events in terms of speaker, listener, and

the interval in which the push-to-talk button was depressed. A Radio Design Lab audio matrix also enabled experimenters to control the status of all lines of communication.

Custom software (seven applications connected over a local area network) ran the synthetic task and collected values of various parameters that were used as input by performance scoring software. A series of tutorials were designed in PowerPoint for training the three team members. Custom software was also developed to conduct tests on information in PowerPoint tutorials, to collect individual and consensus taskwork relatedness ratings, to collect NASA TLX and SART ratings, to administer knowledge questions, and to collect demographic and preference data at the time of debriefing.

In addition to software, some mission-support materials (i.e. rules-at-a-glance for each position, two screen shots per station corresponding to that station's computer displays, and examples of good and bad photos for the PLO) were presented on paper at the appropriate workstations. Other paper materials consisted of the consent forms, debriefing forms, checklists (i.e. set-up, data archiving and skills training), forms for experimenter recording of process, repeat participants forms, knowledge tests (i.e., secondary measures and teamwork) and a leadership survey.

4.2.3 Primary Measures

In this project we apply and refine some measures of team performance, process, and cognition that we have previously developed and evaluated. In this section the primary measures that we use throughout this project are described.

Performance, team process behaviors, and knowledge measures (including knowledge relevant to situation awareness) are the focus of this project. Demographic items, video records, communication records, subjective measures of situation awareness and workload, a leadership survey, and various individual difference variables were also collected. However, they are secondary to the other measures that are the focus of this report and are described in the following section as "secondary measures."

Team performance. Modifications were made to our previous metric of team performance (Cooke, et al., 2001) in order to base team performance on the rate with which tasks were completed (e.g., number of photos per minute) rather than the proportion of tasks that were completed (e.g., number of photos taken out of total possible). This revision accommodates scoring of the high workload scenario, and other variations of the mission scenarios, and prevents penalizing teams for not achieving similar proportions of outcome across different scenarios. For example, the new team performance metric, which is based on rate of performance, does not penalize teams for photographing a smaller proportion of targets in the high workload missions (e.g., 12 out of 20) despite the improvement from the low workload missions (e.g., 9 out of 9).

Furthermore, in order to make the team score more independent from the individual role scores, we removed penalties for fuel, film, and route sequence violations, as these penalties are specific to only one role. Finally, the relative weighting scheme used in the team performance and

individual role performance metrics was also revised to better differentiate between team and individual tasks or components. For example, the "missed or slow photo penalty" component was given lower weight for the PLO score but higher weight for the team score, as this task requires effort on the part of all team members and is not solely the PLO's responsibility. In general, components of the individual role performance metrics were given a higher weight if those components, or tasks, were controlled solely by that role. Appendix B shows the weighting scheme used for each component of the team and individual role performance metrics.

Team performance was measured using a composite score based on the result of mission variables including time each individual spent in an alarm state, time each individual spent in a warning state, rate with which critical waypoints were acquired, and the rate with which targets were successfully photographed. Penalty points for each of these components were weighted *a priori* in accord with importance to the task and subtracted from a maximum score of 1000. Team performance data were collected for each of the seven missions.

Each individual role within a team (AVO, PLO and DEMPC) also had a composite score based on various mission variables including time spent in alarm or warning state as well as variables that were unique to that role. Penalty points for each of the components were weighted *a priori* in accord with importance to the task and subtracted from a maximum score of 1000. The most important components for the AVO were time spent in alarm state and course deviations, for the DEMPC they were critical waypoints missed and route planning errors, and for the PLO, duplicate good photos, time spent in an alarm state, and number of bad photos were the most important components. *Individual performance* data for a role were collected for each of the seven missions.

Team process behavior. Team process behavior was scored independently by each of the two experimenters. For each mission the experimenters observed team behavior and responded to a series of six questions (see Appendices C & D). Three of these items (P3, P4, and P6) concerned team behaviors that did or did not occur at designated event-triggers in each mission (e.g., within five minutes after the end of the mission, the team discusses and assesses their performance). These items were scored with either a 0 (not present) or 1 (present). The other three items (P1, P2, and P5) also assessed team behaviors that did or did not occur at designated event-triggers in each mission, but these items were scored on a scale that ranged from very poor/none (0) to either good (2) for P2 and P5 or very good (3) for P1. The sum of scores on these six items was expressed as a proportion of total possible points (10) for a given mission. This proportion formed the critical incident process score for each team.

Four summary scores for each team were also used to assess team process for a given mission. Summary scores were based on experimenter judgments on four dimensions (communication and coordination, team decision-making, team situation awareness behaviors, and process behaviors), which were scored on a five-point scale that ranged from 1 (terrible) to 5 (excellent). Experimenters were aided when making their judgments by informal tallies that were kept for each dimension throughout the mission. Appendix E contains the description that the experimenters were given for each process dimension.

Team situation awareness. The product of team situation awareness (i.e., fleeting knowledge of the situation) was measured using two SPAM (Situation Present Assessment Method)-like queries (Durso, Hackworth, Truitt, Crutchfield, Nikolic & Manning, 1998) administered at two randomly selected 5-minute intervals during each mission. One of the experimenters administered the queries to each individual in turn (See Appendix F) and then to the team as a whole. Order in which individuals were queried was also random. The two queries asked: (1) a prediction regarding the number of targets out of nine for low workload missions, or 20 for high workload missions, successfully photographed by the end of the mission; and (2) one of the non-repeated queries that is listed in Appendix G. The experimenter also recorded the correct response to these queries once known and this key, which is described below, was used to score the eight responses for accuracy.

Each team member was scored for accuracy at each query (i.e., two accuracy scores per mission). Accurate responses were scored as 1 whereas inaccurate responses were scored as 0. Collective team accuracy on each query was determined by summing the three individual accuracies on that query (i.e., if all 3 individuals were accurate, team accuracy = 3; if 2 individuals were accurate, team accuracy = 2; etc.). Each team was also scored for team holistic team accuracy at each query. Accurate holistic responses were scored as 1 whereas inaccurate holistic responses were scored as 0 and were based on the teams' response to the queries.

Responses to all queries were also scored for intrateam similarity. Team similarity was the sum of all the pairwise similarities of the three team members. First, if the AVO-PLO responses were identical, a score of 1 was assigned to that comparison; otherwise, a score of 0 was assigned. The AVO-DEMPC response and the PLO-DEMPC response were compared and scored in the same way. Intrateam similarity was determined by summing the pairwise comparisons (i.e., if all responses were the same, intrateam similarity = 3; if two team members answered the same, intrateam similarity = 1; if no team members answered the same, intra-team similarity = 0). It was not possible to obtain an intra-team similarity score of 2. In some cases, the truth associated with the query changed as the question was being asked to each of the individuals. That is, on occasion, the situation changed before the experimenter could obtain responses from each individual team member and the team as a whole. In these cases, teams were not scored for intra-team similarity since their answers would have been inaccurate if they were similar and accurate if they were different.

Teamwork knowledge. Team knowledge was measured in two separate sessions by four methods: teamwork questions, teamwork consensus questions, taskwork ratings, and taskwork consensus ratings. Teamwork knowledge was assessed with the teamwork questionnaire (see Appendix H). The teamwork questionnaire consisted of a scenario in which each individual participant was required to indicate which of sixteen specific communications were absolutely necessary in order to achieve the scenario goal. To calculate each individual's overall accuracy, the responses were compared to an answer key, which classified each of the 16 communications into one of the following categories: (1) the communication is NEVER absolutely necessary to complete the scenario goal; (2) the communication could POSSIBLY be necessary to complete the scenario goal (e.g., as considered by novices); or (3) the communication is ALWAYS absolutely necessary to complete the scenario goal. Each communication was worth 2 points, which yielded a maximum of 32 points possible. Participants either checked each

communication, indicating that it was absolutely necessary to complete the scenario goal or left it blank, indicating that it wasn't absolutely necessary. The table below illustrates how the questionnaires were scored. A perfect score was achieved by only checking those communications that were ALWAYS absolutely necessary and leaving all other communications blank.

Table 3
Points Assigned to Responses on the Teamwork Questionnaire

TRUTH	IF PARTICIPANT CHECKED RESPONSE	IF PARTICIPANT LEFT ITEM BLANK
NEVER Necessary	✓0 points given	2 points given
POSSIBLY Necessary	✓1 point given	2 points given
ALWAYS Necessary	✓2 points given	0 points given

Using the same scoring scheme, individual's responses to the teamwork questionnaire were also scored against role-specific keys. In particular, "role" or "positional" accuracy, as well as "interpositional" accuracy (i.e., interpositional knowledge or knowledge of roles other than his or her own) was determined. Role or positional knowledge accuracy was determined by comparing each individual's responses to the role-specific key.

To score positional knowledge accuracy, each role-specific key was used to compare each individual's responses to the subset of the items on the questionnaire specific to his/her role. For example, the key for AVO positional knowledge did not take into consideration five items on the questionnaire that asked about communications between PLO and DEMPC. Therefore, the maximum score for AVO positional knowledge accuracy was 22 (i.e., 11 questionnaire items worth 2 points each). The maximum scores for PLO and DEMPC positional knowledge accuracy were 20 and 22, respectively.

For each role, interpositional knowledge was scored against those items on each key not used in scoring positional knowledge. For example, the accuracy of AVO's responses on the teamwork questionnaire to those 5 items involving communications between the PLO and DEMPC constituted his/her score for interpositional knowledge. Since each response is worth 2 points, the AVO interpositional knowledge maximum is 10. The maximum scores for PLO and DEMPC interpositional knowledge accuracy scores were 12 and 10, respectively.

Intra-team similarity was also computed by comparing responses from all 3 participants and assigning a point to every response that all the team members had in common. A maximum of 16 points were possible where a higher score indicates that the team's responses were more similar.

The teamwork consensus ratings were administered in the same manner as the teamwork ratings, but were completed on a team level where team members discussed their answers over the headsets until a consensus was reached. In this manner, each team was scored for holistic accuracy on the teamwork variable, for a maximum score of 32.

Taskwork knowledge. Taskwork knowledge was assessed through a rating task. The taskwork ratings consisted of eleven task related terms: altitude, focus, zoom, effective radius, ROZ entry,

target, airspeed, shutter speed, fuel, mission time, and photos. These task related terms formed 55 concept pairs, which were presented in one direction only, one pair at a time. Pair order was randomized and order within pairs was counterbalanced across participants.

Team members made relatedness ratings of the 55 concept pairs on a six-point scale that ranged from unrelated to highly- related. By submitting these ratings to Knowledge Network Organization Tool (KNOT), using parameters r = infinity and q = n-1, an individual Pathfinder network (Schvaneveldt, 1990) was derived for each of the team members. These networks reduce and represent the rating data in a meaningful way in terms of a graph structure with concept nodes standing for terms and links standing for associations between terms. The individual taskwork networks were scored not only against a key representing overall knowledge, but also against role-specific keys. In this way, measures of "role" or "positional" accuracy, as well as "interpositional" accuracy could be determined. In previous studies the referent networks were derived manually by experimenters, who were familiar with the UAV-STE. In the experiments presented here we decided that the referents might be improved by basing them on data from the highest scoring individuals or teams in our previous studies. See Appendix I for overall and positional referent networks and the approach that was used to derive these networks.

The accuracy of an individual's knowledge was determined by comparing each individual network to empirical referents associated with knowledge relevant to the respective roles and overall knowledge. Network similarities were computed that ranged from 0 to 1 and represented the proportion of shared links between the two networks (i.e., based on the Pathfinder similarity metric).

From these similarity values, three accuracy values were computed for each team member. Overall accuracy is the similarity between the individual network and the overall knowledge referent. Positional (role) accuracy is the similarity between the individual's network and the referent network associated with that individual's role. Interpositional accuracy is the average of the similarity between the individual's network and the referent networks of the two other roles. These three accuracy values were averaged across all team members to give a final overall, positional and interpositional accuracy score for each team. It should be noted that prior to averaging similarity values to calculate positional and interpositional accuracy scores for the team, positional and interpositional scores for each team member were standardized, as team positional and interpositional accuracy scores are made up of individual scores based on different referents, or scales.

Intrateam similarity was scored on the same scale as accuracy and ranged from 0 to 1. An individual's network was compared to another team member's network and assigned a similarity value. This was done until all three team members had been compared to one another (i.e. AVO-PLO, AVO-DEMPC, and PLO-DEMPC). Intrateam similarity was computed by averaging the three similarities values and measured using the proportion of shared links for all intrateam pairs of two individual networks (i.e. the mean of the three pairwise similarity values among the three networks).

Taskwork consensus ratings consisted of the same pairs as taskwork ratings (randomly presented); however the *team* entered a rating for each pair. For each pair, the rating entered in the prior session by each team member was displayed on the computer screen of that team member. The three team members discussed each pair over their headsets until consensus was reached. As a team, the individuals had to agree on relatedness ratings for the concepts. The team ratings were submitted to Pathfinder network scaling. The holistic accuracy score is the similarity value between the team's network and the overall referent network. From their answers, a team knowledge network was developed and compared to the overall knowledge referent.

4.2.4 Secondary Measures

Verbal working memory capacity. There were several secondary measures that will be briefly described in this section. A measure of verbal working memory capacity was taken from the Air Force CAM 4 computerized test battery (Kyllonen, 1995; Kyllonen & Christal, 1990). This measure consisted of 32 items, each of which presented participants with four to seven stimuli, the last three of which the participants had to remember in order. The stimuli were one-syllable adjectives such as big, cold, and fast. The color of the stimuli was varied so that participants had to transform the words. When the stimulus was white, the participant remembered the word that had been presented, but when the word was yellow, participants had to remember the antonym of the word (e.g., the opposite of big is small). The last three stimuli for an item were either consistent, that is all white or all yellow, or were inconsistent, which means that white and yellow stimuli were mixed. Stimuli were presented at the rate of one word every 2.5 seconds and participants had 18 seconds to respond to each list of words.

Participants responded after all stimuli had been presented by selecting from eight alternatives: big, cold, fast, high, hot, low, slow, and small. Participants also used a three-point scale to provide confidence ratings concerning the correctness of their responses. The three points on the scale were labeled "I think I got it wrong," "I am not sure", and "I think I got it right". After each block of eight trials, participants were provided with accuracy feedback. The feedback also indicated how many items the participant thought he/she had answered correctly. The working memory task presumably tapped working memory capacity because the participants were required to retain and manipulate the stimuli. Results pertaining to this measure are described in the section on archival analyses of individual and role-related factors.

Social desirability. Another individual difference measure that was used in Experiment 1 was a computerized version of the Marlowe-Crowne social-desirability scale (MCSD; Crowne & Marlowe, 1964). This measure asked participants to respond to 33 true-false items as part of the debriefing questions (see below). Higher scores on this measure may indicate a need for approval and a certain amount of defensiveness. Due to lack of variance on this measure, results pertaining to it are not reported.

Verbal processing speed. Another individual difference measure that was used in Experiment 2 only was a verbal processing speed measure that assessed how quickly the participant could decide whether simple words had similar meanings or different meaning (e.g., argue and debate). Participants responded by typing "L" on the computer if the words were synonyms and "D" if

not. Participants received latency feedback at the end of each trial and accuracy feedback at the end of each block. The measure of performance was response latency, which is referred to as processing speed in this report. Results pertaining to this measure are described in the section on archival analyses of individual and role-related factors.

Secondary knowledge questionnaire. A secondary questionnaire was also administered at each of the two knowledge sessions in Experiment 1. This 20-item questionnaire provided a secondary measure of taskwork and teamwork knowledge that was necessary for planned MTMM analyses. For 16 items, participants used a five-point scale to provide ratings of the knowledge and abilities of (1) themselves, (2) their teammates, and (3) the team as a whole. The remaining 4 multiple-choice items directly assessed the participants' knowledge and abilities. The secondary questionnaire can be found in Appendix J. Results pertaining to this measure are reported in the section on archival evaluation of measures, specifically in the section on MTMM analysis.

Emerging leadership. In the first experiment, a leadership questionnaire was administered in order to attempt to assess emerging leadership. A copy of this questionnaire appears in Appendix K. Results pertaining to this questionnaire are reported separately as part of Rebecca Keith's Masters thesis at NMSU.

NASA TLX subjective workload measure. Another questionnaire was administered at the end of each mission. This post-mission questionnaire included two measures. The first measure was a variant of the NASA TLX, a subjective measure of workload (Hart & Staveland, 1988). This subjective workload measure asked participants to respond to 5 different rating scales (i.e., mental demands, physical demands, temporal demands, performance demands, teammate demands) regarding their last mission. Each scale contained one question and the responses were on a scale from 1 (low demands) to 100 (high demands). These numbers were not shown to the participants. The ratings on each subscale were weighted according to the extent to which each type of demand contributes to the workload in our task (as decided by experimenters). For example, our task requires more mental demand (remembering, deciding, etc.) than physical demand (pushing, pulling, etc.) and thus, mental demand is weighted more heavily than physical demand. Furthermore, these weights differ among the roles, as each type of demand does not necessarily contribute to each role's workload in the same manner (see Table 4). Results pertaining to this measure are described in the section on archival analyses of individual and role-related factors and can also be found in the appendix that pertains to workload measures.

Table 4
The Role-Specific Weights for each Subscale on the NASA TLX

	Mental	Physical	Temporal	Performance	Team
AVO	1.67	1.17	2.67	2.33	2.17
PLO	2.00	.67	2.50	3.00	1.83
DEMPC	3.67	.33	1.17	1.50	3.33

SART: Subjective situation awareness measure. The second measure on the post-mission questionnaire was the SART (Situation Awareness Rating Technique; Taylor, 1990), which is a subjective measure of situation awareness. The subjective situation awareness questionnaire was

made up of 14 ratings (e.g., demand on cognitive resources, instability of situations, complexity of situations, etc.), each on a scale from 1 (low) to 7 (high). The post-mission questionnaire containing the subjective measures of both workload and situation awareness can be found in Appendix L. Results for this measure are presented with the results for situation awareness.

Debriefing questions. We also administered a series of questions at the end of the study to assess various constructs such as trust and evaluation anxiety as well as collect demographic information. A set of questions also asked participants about their experiences as a participant such as whether they enjoyed the study, liked working with other members of the team, performed well on the task, and how they felt about other members of their team. Participants were also asked about how they made decisions (e.g., majority-rules, unanimous). The complete set of questions for each of the three studies can be found in Appendices M and N for Experiments 1 and 2, respectively. The debriefing in Experiment 3 involved a demographics questionnaire and a debriefing interview (see Appendices O and P, respectively). Due to the unique background of one of the teams in Experiment 3, eight additional questions were appended to the Debriefing Interview (see Appendix Q). Results pertaining to this measure are described in the section on archival analyses of individual and role-related factors.

4.2.5 Procedure

Experiment 1 consisted of two sessions (see Table 5). Both sessions lasted approximately 5 hours each and were separated by a 48-hour interval. Teams were randomly assigned to one of two conditions (distributed or co-located). In the distributed condition, two team members were in the same room separated by partitions and the third team member was in a remote location. In the co-located condition, all three team members were in the same room. Prior to arriving at the first session, the three participants were randomly assigned to one of the three task positions: AVO, PLO, or DEMPC. The team members retained these positions within the same team for the remainder of the study.

Table 5
Experiment 1 Protocol

SESSION 1	*SESSION 2*
Working Memory Measure	Mission 4 (low workload)
Task Training	Post Mission Questionnaire
Knowledge Measures	Leadership Questionnaire
Mission 1 (low workload)	Mission 5 (high workload)
Post Mission Questionnaire	Post Mission Questionnaire
Mission 2 (low workload)	Mission 6 (high workload and communication glitch)
Post Mission Questionnaire	Post Mission Questionnaire
Mission 3 (low workload)	Mission 7 (high workload)
Post Mission Questionnaire	Knowledge Measures
	Leadership Questionnaire
	Debriefing Questions

In the first session, the team members were seated at their workstations where they completed a working memory measure. Afterwards, they were given a brief overview of the study and started training on the task. During training, all the team members were separated by partitions regardless of the condition they were assigned. Team members studied three PowerPoint training modules at their own pace and were tested with a set of multiple-choice questions at the end of each module. If responses were incorrect, they were instructed to go back to the PowerPoint tutorial and correct their answers. Experimenters provided assistance and explanation if their second response was also incorrect. Once all team members completed the tutorial and test questions, a mission was started and experimenters had participants practice the task, checking off skills that were mastered (e.g., the AVO needed to change altitude and airspeed, the PLO needed to take a good photo of a target) until all skills were mastered (See Appendix R for the checklist of skills). Again, the experimenters assisted in cases of difficulty. Training took a total of 1.5 hours.

After a short break, knowledge measures were administered in the following order: taskwork ratings, taskwork consensus ratings, teamwork ratings, teamwork consensus ratings, and the secondary knowledge questionnaire. The participants were separated by partitions during the knowledge sessions as well. Once the knowledge measures were completed, partitions were removed for co-located teams and the teams began the first 40-minute mission. Missions 1 through 4 were low workload, which required teams to take reconnaissance photos of 9 targets. Missions were completed either at the end of a 40-minute interval or when team members believed that the mission goals had been completed. Immediately after each mission, participants were shown their performance scores. In the co-located condition, participants could view their team score, their individual score, and the individual scores of their teammates. Participants in the distributed condition were only allowed to view the team score and their own individual score. The performance scores were displayed on each participant's computer and shown in comparison to the mean scores achieved by all other teams (or roles) who had participated in the experiment up to that point. Teams also completed a set of post mission questions following each mission, which included the SART and TLX that were described earlier. Participants were given their second break after Mission 1.

The second session consisted of Mission 4 followed by a short leadership questionnaire. The last three missions were high workload and required participants to take reconnaissance photos of 20 targets. There were also additional constraints relevant to route planning. The participants took a break following Mission 5. During Missions 6 and prior to entry by the team's UAV into a specific target area, communication was cut for five minutes so the AVO could not hear any information relayed by the DEMPC. Mission 7 was then completed followed by a break and the second knowledge measurement session. During the second knowledge session, participants completed the same ratings as in the first knowledge session, as well as leadership and debriefing questionnaires.

4.3 Experiment 1: Results

This section describes work under Task 2 of the first objective: *analyze data from the first experiment to determine the direction for the follow-up experiment.* As stated earlier, team performance, team process behaviors, and knowledge measures (including knowledge relevant to

situation awareness) are the focus of this project and are reported in the results section that follows. Results are summarized at the end of each section to facilitate an understanding of the main points. Some detailed analyses of workload measures are presented in the appendix (see Appendix S).

4.3.1 Team performance

Table 6 shows the means for the co-located and distributed teams for each mission and Figure 5 provides a graph of these means. A mixed two-factor ANOVA with mission as the repeated measure and dispersion as the between-teams variable revealed a detectable effect of mission $F(6, 108) = 19.10, p < .01$, but no effect of dispersion $F(1, 18) < 1$.

Table 6
Team Performance in Co-located and Distributed Conditions

Mission	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	290	239	79	58	163	162	397	336
2	340	327	116	42	63	268	481	389
3	378	398	54	47	312	337	472	479
4	446	420	73	51	317	362	539	495
5	338	378	40	53	270	327	385	467
6	355	360	44	34	269	303	407	430
7	357	392	56	45	247	327	462	457

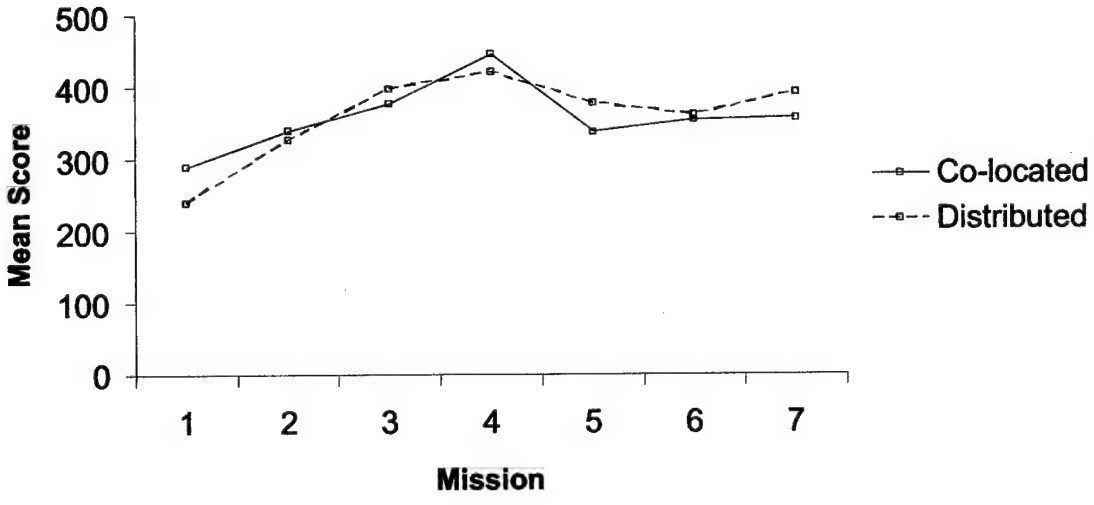


Figure 5. Performance scores for co-located and distributed teams.

Sequential acquisition contrast effects are shown in Table 7. There was also an interaction between mission and dispersion $F(6, 108) = 1.94$, $p = .08$. (Note that throughout this report we considered α-levels of $\leq .10$ statistically significant, opting to err in the direction of increased Type I errors in order to identify any potentially interesting measures or effects.) Simple effect contrasts showing the difference between dispersion levels between each mission from 2 through 7 are presented in Table 8. The comparison at Mission 1 is excluded due to degrees of freedom constraints.

Table 7
Sequential Acquisition Contrast Effects for Performance-Means are Adjusted for the Repeated Measures Model

Contrast Between Missions	*B* (mean difference)	SE_B	β	*t*	*p*
2 - 1	94.09	10.99	.66	8.56	<.01
3 - 2	119.21	14.19	.84	8.40	<.01
4 - 3	89.75	15.55	.63	5.77	<.01
5 - 4	15.34	15.55	.11	0.99	.33
6 - 5	15.39	14.19	.11	1.08	.28
7 - 6	16.33	10.99	.11	1.49	.14

Table 8
Performance Dispersion Effects at Mission 2 through 7-Mission 1 Excluded to Preserve Degrees of Freedom

Dispersion at Mission:	*B* (mean difference)	SE_B	β	*t*	*p*
2	-18.92	16.79	-.09	-1.13	.26
3	-35.45	16.79	-.18	-2.11	.04
4	-12.58	16.79	-.06	-0.75	.46
5	-45.22	16.79	-.22	-2.69	.08
6	-27.73	16.79	-.14	-1.65	.10
7	-42.53	16.79	-.21	-2.53	.01

Improved performance during the first four missions was tested by comparing the means from Missions 1 and 4. Analysis with Missions 1 and 4 as the repeated measure and dispersion as the between-teams variable indicated that teams in both conditions learned the task. There was no interaction between condition and mission performance $F(1, 18) < 1$, nor main effect of condition $F(1, 18) = 2.25$, but performance improved between Missions 1 and 4 $F(1, 18) = 124.46$, $p < .01$, averaged across co-located and distributed teams.

Comparisons were also made between performance in Mission 4 and the Mission 4 performance of teams from an earlier study (Cooke, et al., 2001) to determine whether the teams reached asymptote in Mission 4 as they had in the earlier study. A two degree of freedom test including co-located teams against earlier teams, and distributed teams against earlier teams, produced no detectable difference, $F(2, 28) < 1$, in Mission 4 performance, so we assume that both co-located and distributed teams in the present study reached asymptote in Mission 4.

Increased workload produced a decline in performance between the last low workload mission (Missions 4) and the first high workload mission (Mission 5), $F(1, 18) = 31.47, p < .01$, (see Table 6 for means and SDs), with a detectable interaction between dispersion condition and mission $F(1, 18) = 6.05, p = .02$, suggesting that the decline in performance was affected by condition. Means in this single degree of freedom interaction reveal that the direction of the dispersion effect changed from Mission 4 to Mission 5, with distributed teams performing better than co-located in Mission 5 and with co-located teams suffering the most from increased workload (see Table 6 for means and SDs).

We also compared co-located and distributed teams on Missions 4, the last low workload mission, and Mission 7, the last high workload mission, to see whether teams recovered from the workload manipulation by the end of the experiment. There was a detectable main effect of mission $F(1, 18) = 13.74, p < .01$, with teams performing worse in Mission 7 than in Mission 4 (see Table 6 for means and SDs). Also, there was an interaction between condition and mission, $F(1, 18) = 3.65, p = .07$, indicating a change in valence for the dispersion effect between Mission 4 and Mission 7 with distributed teams outperforming co-located teams in Mission 7, but not in Mission 4.

To summarize:

- Co-located and distributed teams learned the task during the low workload missions and performed more poorly when the workload was increased.

- Dispersion did not affect early performance, but the dispersed teams tended to perform better than the co-located teams on later high workload trials.

In sum our hypotheses regarding performance deficits of DME teams were not supported by our findings. In fact, the deficits seem to point to the co-located teams.

4.3.2 Team Process

To calculate agreement between the two process raters, we computed a scaled proportion of agreement index (Po(scale); Cooke, et al., 2001). Between all pairs of raters, we computed the absolute value of the deviation, scaled to the range of the possible scores. This normalized disagreement measure was then subtracted from 1, yielding:

Po(scale) = 1 - |Rater1 - Rater2| / Range.

Next, for each mission, we tested the Po(scale) for each process measure using a one-sample t-test, against 0. Every process measure at every mission (both critical incident and rating measures) was detectably larger than disagreement, with no p-value being larger than .01, and almost all of them being smaller than .001 (see Appendix Table T1). Therefore, agreement was adequate for the process measures, and we averaged between the two raters to yield an overall score for each item.

Critical incident process. As discussed in the measures section, the sum of scores on the six critical incident items was expressed as a proportion of total possible points (10) for these items in a given mission. Thus critical incident process ranges from 0 to 1. The items were equally weighted in the proportion based on the results of a hierarchical centroid clustering of the items. For the distance metric we subtracted the rank order correlations from 1. This metric ranges from 0 to 2, with 0 being closest and 2 being farthest. Depicted in Figure 6, the distance between each cluster is roughly linear, with no dramatic jumps, we therefore felt confident there were no strong clusters among the individual items, with each item being equally important in the overall score.

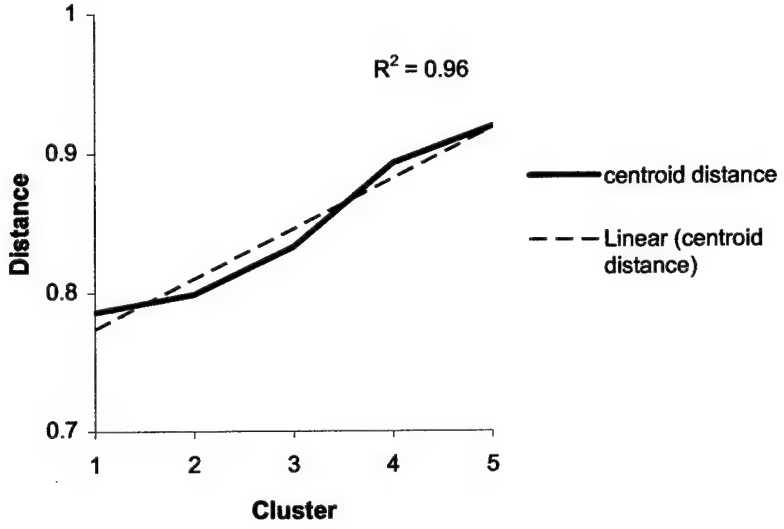


Figure 6. Experiment 1 critical incident process items: distance by cluster.

When the team did not reach a designated event-trigger, and therefore had missing data for that item, the proportion was calculated without that item. In this way, the team's interaction score was not affected by an event that is better captured by the team's performance score. Descriptive statistics including means, standard deviations and ranges for critical incident process scores for co-located and distributed teams at each mission are given in Table 9

Table 9
Team Critical Incident Process Scores in Co-located and Distributed Conditions

Mission	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	.56	.35	.15	.13	.35	.20	.80	.60
2	.66	.51	.10	.08	.50	.40	.80	.65
3	.70	.55	.15	.17	.45	.35	.85	.80
4	.66	.48	.08	.13	.50	.25	.75	.65
5	.56	.50	.16	.16	.40	.22	.80	.75
6	.64	.49	.11	.12	.50	.35	.85	.75
7	.66	.48	.14	.09	.39	.35	.90	.60

Based on our experimental design, we analyzed critical incident process for main effects of mission and condition, and the interaction mission by condition. The levels of the factors are 2 conditions by 7 missions, with 10 teams per condition, N = 2 X 10 X 7 = 140 total observations. Planned comparisons included looking at interactions between condition and Missions 1 and 4, as well condition and Missions 4 and 5. The first contrast tests for differences in acquisition or learning while the second tests for differential workload effects.

The main effect of mission was significant, $F(6, 108) = 4.77$, $p < .01$, implying that critical incident process scores were statistically different across the seven missions. The main effect of condition was also significant, $F(1, 18) = 18.41$, $p < .01$. Figure 7 illustrates that co-located teams had significantly higher critical incident process over missions. Figure 7 supports the statistical finding that the omnibus interaction between condition and mission was insignificant $F(6, 108) < 1$. Therefore, over all of the missions, differences between co-located and distributed critical incident process were relatively consistent.

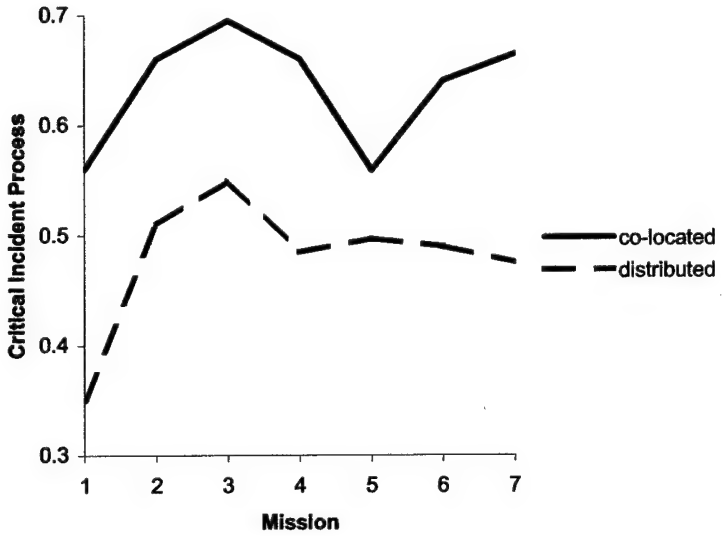


Figure 7. Mean Experiment 1 co-located and distributed critical incident process scores over missions.

The planned comparison between Missions 1 and 4 revealed a main effect of condition, $F(1, 18) = 21.65$, $p < .01$. However this is not surprising since co-located critical incident process was always higher. The main effect of mission for the planned comparison was also significant, $F(1, 18) = 9.94$, $p < .01$, indicating that critical incident process scores statistically increased between Missions 1 and 4. This difference was also consistent across conditions, as the planned interaction contrast between these two levels of mission was not significant, $F(1, 18) < 1$. According to this set of planned comparisons, co-located teams had higher critical incident process scores at both Missions 1 and 4, but that teams in both conditions showed increases in critical incident process scores between Missions 1 and 4.

The planned comparisons between Missions 4 and 5 indicate a significant main effect of condition, $F(1, 18) = 5.53, p < .05$, between these missions, with co-located continuing to earn higher critical incident process scores. The main effect of mission however was not significant, $F(1, 18) = 1.86$. However inspecting Figure 7, the reason this main effect was not significant is likely due to the fact that the process of distributed teams was not impaired by the higher workload. A *post hoc* simple comparison revealed that co-located teams did show significant decrease on critical incident process between Missions 4 and 5, $F(1, 9) = 4.37, p = .07$. It is not surprising then that the planned interaction contrast using Missions 4 and 5 was also significant, $F(1, 18) = 3.01, p = .10$. Apparently co-located and distributed teams' critical incident process scores were differentially impacted by the transition from low to high workload between Missions 4 and 5. In Figure 7 this difference is illustrated by the sharp drop in co-located critical incident process at Mission 5 relative to the steady, albeit low, level of critical incident process at Missions 4 and 5 for distributed teams.

Reflecting on these results, it appears that team process behaviors change over time. Looking at Figure 7, we can see several bumps and dips across missions. On the positive side this implies that team process behaviors can be made to be adaptive, in so much as they are dynamic and can change over time. On the negative side, this also implies that process behaviors can be relatively unstable, requiring a high level of maintenance. Our results also suggest that teams in the co-located condition exhibit better team process behaviors at our pre-defined trigger points. Given that our participants in the co-located condition are geographically proximal, the simple explanation is that something about being together in the same room facilitates these good process behaviors. And although the co-located teams took a bigger hit to their process behaviors in the high workload missions, they were still far better than their distributed counterparts.

Based on the results from the planned analyses of critical incident process, we conducted some follow up tests. In order to more deeply explore the source of the process differences between co-located and distributed teams, a discriminant analysis model was fit using the critical incident items as predictors and co-located (0) or distributed (1) as the dependent grouping variable. Wilks' Lambda and the *F* analogue of the weights assigned to each item in the discriminant function are presented in Table 10.

Table 10
Results of Discriminant Analysis

Process Item	Wilks' Lambda	*F*	*df*num	*df*den	Sig.	Standardized Weights
1	.90	13.31	1	124	.00	.16
2	1.00	.00	1	124	.99	-.11
3	.94	7.38	1	124	.01	.28
4	1.00	.50	1	124	.48	-.06
5	1.00	.01	1	124	.92	-.11
6	.34	244.84	1	124	.00	.98

Clearly critical incident Item 6 is the big discriminator, followed by Items 1 and 3 in that order. It is interesting to note that these items involve communications that are not explicitly necessary.

For example, Items 6 and 1 involved teams discussing their performance after and at the beginning of, respectively, their missions. Item 3 is whether or not they explicitly note call in targets before getting to a called in ROZ (Restricted Operating Zone) area. All of the other Items, 2, 4, and 5, involve communications that are explicitly necessary during the course of a mission, e.g., AVO and PLO coordinating on a specific target. We thus theorize that the significantly better process behaviors exhibited by co-located teams were due to differences in assessing performance prior to and after each mission (Items 1 and 6), and to some extent, explicitly noting mission parameters that emerge during the course of a mission (Item 3). Although these differences apparently do not correspond to differences in team performance, these items may be relevant in terms of planning and adaptive process behaviors.

Summary process. Based on the previous inter-rater agreement for process measures, each of the four summary process ratings were averaged between the two experimenters at that mission. If a value was missing for one of the experimenters the available value was taken as the team's mission score on that process dimension.

In order to compute an overall summary process score we averaged over the four components. The average was deemed appropriate based on hierarchical centroid clustering of the 1-rho distance metric. Specifically, as illustrated in Figure 8, there were no obvious clusters among subsets of summary process dimensions. Descriptive statistics of this overall summary process measure for co-located and distributed teams over missions are presented in Table 11.

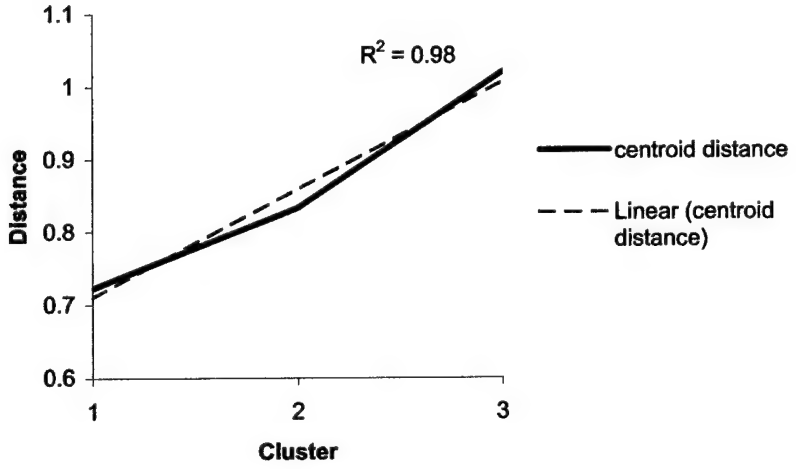


Figure 8. Experiment 1 summary process items: distance by cluster.

Table 11
Team Process Summary Scores in Co-located and Distributed Conditions

Mission	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	2.56	2.22	.94	.45	1.63	1.50	4.50	3.00
2	2.76	3.19	.76	.43	1.63	2.38	4.00	3.75
3	3.41	3.42	.83	.51	2.00	2.75	4.75	4.25
4	3.81	3.71	.43	.66	2.88	2.50	4.38	4.75
5	3.25	3.55	.63	.63	2.50	2.88	4.25	4.88
6	3.58	3.51	.42	.43	2.75	2.88	4.13	4.38
7	3.41	3.50	.73	.63	2.00	2.50	4.13	4.50

The planned analyses of the averaged summary process ratings were the same as those for critical incident process. Specifically we tested for main effects of mission and condition and an interaction effect between the two, followed by interaction contrasts involving co-located and distributed differences in acquisition (condition by Missions 1 and 4) and workload (condition by Missions 4 and 5). There were 140 total observations.

The main effect of mission was significant, $F(6, 108) = 13.76$, $p < .01$, indicating that summary process scores changed across missions. In Figure 9, it appears that in general, summary process increased across the first four (low workload) missions, but dipped during the last three (high workload) missions. Unlike critical incident process, the main effect of condition on summary process was not significant, $F(1, 18) < 1$. Apparently the distributed teams' poorer critical incident process did not reflect on experimenter ratings of quality of team process behaviors. The interaction between mission and condition was also not significant, $F(6, 108) = 1.08$. Thus, as Figure 9 suggests, the differences in summary process across missions were consistent between co-located and distributed conditions. The planned comparisons supported this general finding between specific levels of the mission factor.

Between Missions 1 and 4 there was a large increase in mean summary process ratings, $F(1, 18) = 51.27$, $p < .01$, that was consistent across both conditions, $F(1, 18) < 1$. Likewise the dip between Missions 4 and 5 was significant, $F(1, 18) = 6.24$, $p < .05$, and apparently occurred similarly for both conditions, $F(1, 18) < 1$. Both the acquisition (mission 1 vs. 4) and workload (mission 4 vs. 5) interaction contrasts were not significant, $F(1, 18) < 1$ and $F(1, 18) = 1.91$, respectively, and support the conclusion that summary process increased and decreased similarly for co-located and distributed teams.

The results for summary process suggest that the quality of team processes as rated by expert observers were roughly equal between the co-located and distributed conditions. This finding, taken together with the results from the follow up analysis on critical incident process, suggest that the critical incident items on which co-located teams exhibited an advantage may not have been an important factor when considering overall quality of a teams process behaviors. More specifically, in so much as the critical incident items on which co-located teams held an advantage involve planning and adaptive process behaviors, these factors where un-represented among the summary process dimensions. However, the lack of strong differences in team

performance between the co-located and distributed conditions further bolsters the claim that these dimensions may not be necessary in order to evaluate team process behaviors as they relate to performance in the UAV synthetic task. Finally, the summary process results suggest that over the first four missions, the ratings of overall quality of team process behaviors improved. We theorize that this maps onto the performance acquisition curve, in that over the first four missions part of what teams are really acquiring is the ability to coordinate with each other via good process behaviors.

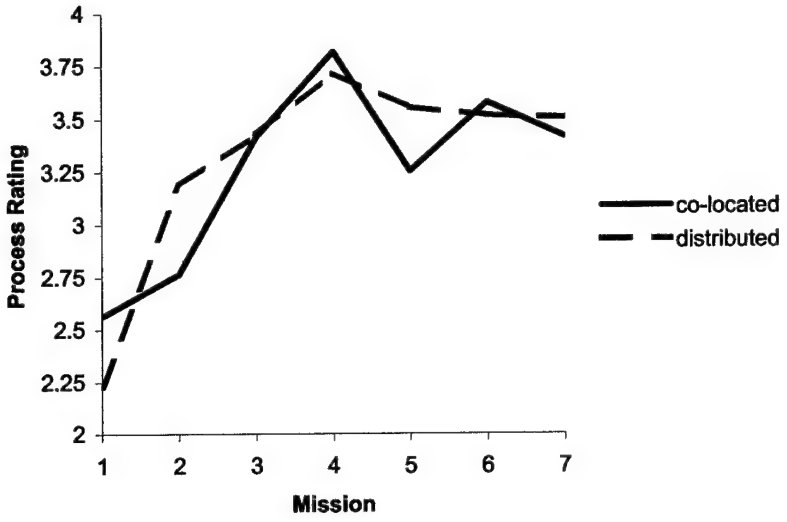


Figure 9. Mean Experiment 1 co-located and distributed summary process ratings scores over missions.

To summarize the most interesting findings:

- Co-located teams exhibited better overall critical incident process behaviors; these behaviors tended to involve planning and adaptive process behaviors, which do not map directly onto team performance differences in the UAV synthetic task.
- At least part of what all teams acquire while approaching performance asymptote is the ability to coordinate using good process behaviors.
- Increases in workload tend to impair team process behavior.

4.3.3. Situation Awareness

The analyses of each of the measures of situation awareness (accuracy, intrateam similarity, and holistic accuracy) examined the effects of mission, condition (co-located/distributed), and type of query (repeated/non-repeated). Of the eight situation awareness queries, one query was repeated at each mission while each of the other (non-repeated) queries was administered at a different mission (determined randomly). Combining repeated and non-repeated queries would not be appropriate, given that there appears to be differences in what the queries measure. That is, the

repeated query seems to measure awareness of the experimental situation since it is administered at each mission and can therefore be anticipated by the team with mission experience. On the other hand, the non-repeated queries cannot be anticipated and act as a measure of awareness of the task situation.

Situation awareness accuracy. Table 12 shows situation awareness accuracy on the repeated query and non-repeated queries for co-located and distributed teams on a mission-by-mission basis as well as averaged over low workload missions and high workload missions. A missing data point for Team 12 at Mission 4 (repeated query) was replaced with the mean of the other 19 teams' accuracy scores on the repeated query at Mission 4 before the overall mean was calculated.

A univariate, repeated measures analysis of variance with two repeated factors (mission and query type) and one between-subjects factor (condition) was used to analyze situation awareness accuracy. Results from the omnibus test are presented first, followed by the results from two planned contrasts. Accuracy for co-located and distributed teams was not significantly different, $F(1, 18) = 2.53$. However, a significant effect of mission emerged, $F(6, 108) = 3.14, p < .01$. Figures 10 and 11 illustrate how accuracy changed across missions. Query type also significantly affected accuracy, $F(1, 18) = 105.11, p < .01$. In particular, teams were more accurate at responding to the non-repeated queries than the repeated query. Furthermore, there was a significant interaction between mission and query type, $F(6, 108) = 6.41, p < .01$, which can be seen in the comparison of Figure 10 and Figure 11. That is, the pattern of accuracy across the missions was different for the repeated query and the non-repeated queries. The interaction between condition and query type was not significant, $F(1,18) < 1$. There was also no interaction between condition and mission, $F(6, 108) < 1$. Finally, the three-way interaction among condition, query type, and mission was not significant, $F(6, 108) < 1$.

Post hoc comparisons were conducted to pin-point the source of the interaction between mission and query type. The comparisons revealed that accuracy on the non-repeated queries was significantly higher than accuracy on the repeated query during Missions 1-2 and Missions 5-7 (see Table 13 for t statistics and p-values).

Table 12

Situation Awareness Accuracy on the Repeated Query and Non-Repeated Queries for Co-located and Distributed Teams

Mission	Mean		St. Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Repeated Query								
1 (LW)	.70	.30	.82	.48	.00	.00	2.00	1.00
2 (LW)	.80	.90	.92	.88	.00	.00	3.00	2.00
3 (LW)	1.30	1.40	1.25	1.17	.00	.00	3.00	3.00
4 (LW)	1.97*	1.40	1.16	1.27	.00	.00	3.00	3.00
5 (HW)	.20	.10	.42	.32	.00	.00	1.00	1.00
6 (HW)	.30	.10	.48	.32	.00	.00	1.00	1.00
7 (HW)	.90	.60	.57	.70	.00	.00	2.00	2.00
Average of Low Workload Missions	1.18	1.00	.75	.54	.00	.25	2.75	2.00
Average of High Workload Missions	.47	.27	.23	.21	.00	.00	.67	.67
Non-Repeated Query								
1 (LW)	1.30	1.70	1.06	.95	.00	1.00	3.00	3.00
2 (LW)	1.90	2.10	1.10	.74	.00	1.00	3.00	3.00
3 (LW)	1.90	1.50	1.29	1.18	.00	.00	3.00	3.00
4 (LW)	1.80	1.80	1.03	.92	1.00	1.00	3.00	3.00
5 (HW)	2.30	2.00	.82	.82	1.00	1.00	3.00	3.00
6 (HW)	2.50	2.10	.53	.99	2.00	.00	3.00	3.00
7 (HW)	2.20	2.10	.79	1.10	1.00	.00	3.00	3.00
Average of Low Workload Missions	1.73	1.78	.49	.42	1.25	1.25	2.50	2.50
Average of High Workload Missions	2.33	2.07	.42	.34	1.67	1.67	2.67	2.67

* Contained missing data for one team, which was replaced with the mission mean

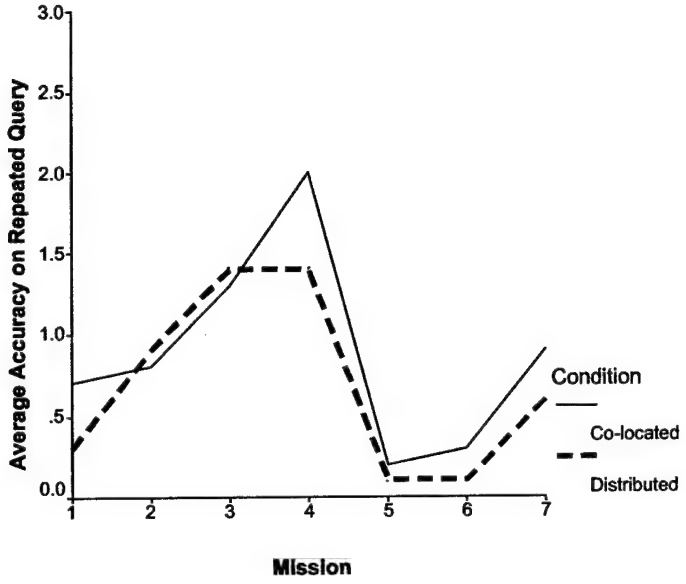


Figure 10. Situation awareness accuracy on the repeated query for co-located and distributed teams at each mission.

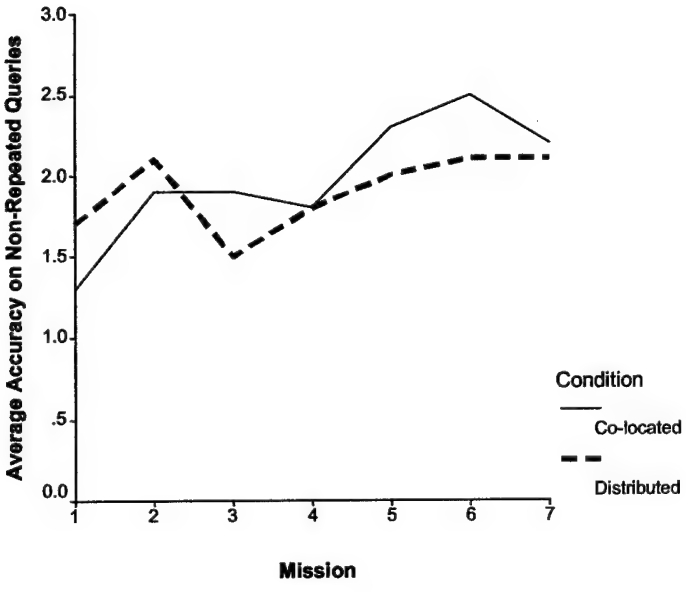


Figure 11. Situation awareness accuracy on the non-repeated queries for co-located and distributed teams at each mission.

Table 13

T Statistics for the Comparison of the Average Accuracy on the Repeated Query Minus Average Accuracy on the Non-repeated Queries at Each Mission

Mission	*t* statistic	*p*-value
1	-3.68	.00
2	-4.06	.00
3	-.92	.36
4	-.34	.74
5	-10.03	.00
6	-10.43	.00
7	-5.54	.00

$df = 38$

A series of planned contrasts were also conducted in order to answer the following questions: (1) Did teams' accuracy improve over the low workload missions on the repeated and non-repeated queries, and 2) was there an effect of workload on accuracy for the repeated and non-repeated queries? Although co-located/distributed status was an initial variable of interest, it was excluded from the following contrasts due to the lack of effects it produced in the omnibus test presented above. A univariate, repeated measures analysis of variance with two repeated factors (mission and query type) was used to analyze each contrast.

First, a comparison of Mission 1 to Mission 4 was used to determine if accuracy improved over the low workload missions. Accuracy did improve significantly from Mission 1 to Mission 4 $F(1, 19) = 19.56, p < .01$. In addition, accuracy was significantly higher on the non-repeated, queries than on the repeated query, $F(1, 19) = 3.50, p = .08$. There was also a significant interaction between mission and query type, $F(1, 19) = 4.70, p = .04$. *Post hoc* comparisons revealed that the mission by query type interaction originated from the fact that accuracy on the repeated query significantly improved from Mission 1 to Mission 4, $F(1, 19) = 19.89, p < .01$, while accuracy on the non-repeated queries did not significantly differ from Mission 1 to Mission 4, $F(1, 19) = 1.31$.

For the second planned contrast, was there an effect of workload on accuracy? In order to test for the effect of workload, accuracy scores from Mission 4 (the last low workload mission) were compared to accuracy score from Mission 5 (the first high workload mission). Mission 5 accuracy was used as the measure of high workload to allow for comparisons to be made between the current experiment and Experiment 2, which only had a single high workload mission (i.e., Mission 5). Accuracy significantly declined from Mission 4 to Mission 5, $F(1, 19) = 12.86, p < .01$. That is, teams were more accurate during the final low workload mission than during the first high workload mission. Teams were also more accurate on the non-repeated queries than on the repeated query, $F(1, 19) = 29.75, p < .01$. Finally, an interaction was found between workload and query type, $F(1, 19) = 11.92, p < .01$. *Post-hoc* comparisons were conducted on Mission 4 and Mission 5 for the repeated and non-repeated queries separately in order to locate the source of the interaction. Comparisons indicated that accuracy on the repeated query significantly decreased in Mission 5, the first high workload mission, $F(1, 19) = 25.73, p < .01$, whereas accuracy on the non-repeated queries did not differ between the low workload mission and high workload mission, $F(1, 19) = 1.09$.

Situation awareness intrateam similarity. Table 14 shows situation awareness intrateam *similarity* on the repeated query and non-repeated queries for co-located and distributed teams on a mission-by-mission basis as well as averaged over low workload missions and high workload missions.

Recall that the truth of the non-repeated situation awareness queries often changed in the midst of administering the queries. Consequently, team members necessarily had to respond differently from one another in order to be accurate. In these cases, a similarity score was not calculated. Instead, each missing data point was replaced with the mean of the mission from which it was missing. Of the 140 total missions (20 teams each with 7 missions), missing data were replaced for 13 missions (denoted in Table 14). For the repeated query, it was not possible for the truth to change in the midst of administering the query because the truth was a value that was determined post-mission. However, for other reasons, two data points at Mission 4 were missing from the intrateam similarity scores on the repeated query. Mission means were also used to replace these missing data.

A univariate, repeated measures analysis of variance with two repeated factors (mission and query type) and one between-subjects factor (condition) was used to analyze situation awareness intrateam similarity. A significant effect of condition was not present, $F(1,18) < 1$. However, the effect of mission was significant, $F(6, 108) = 4.60, p < .01$. As Figures 12 and 13 illustrate, intrateam similarity changed as a function of mission. Furthermore, a main effect of query type was found, $F(1, 18) = 43.57, p < .01$, where teams were more similar in their responses to the non-repeated queries than to the repeated query. There was also a significant interaction between mission and query type, $F(6, 108) = 3.81, p < .01$. No interaction emerged between condition and query type, $F(1, 18) = 1.19$, or between mission and condition, $F(6, 108) < 1$. The three-way interaction among condition, mission, and query type was also not significant, $F(6, 108) < 1$.

Table 14

Situation Awareness Intrateam Similarity on the Repeated Query and Non-Repeated Queries for Co-located and Distributed Teams

Mission	Mean		St. Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Repeated Query								
1 (LW)	.50	.30	.53	.48	.00	.00	1.00	1.00
2 (LW)	1.00	.50	1.15	.53	.00	.00	3.00	1.00
3 (LW)	1.70	1.40	1.16	1.17	.00	.00	3.00	3.00
4 (LW)	2.10	1.90**	.99	1.29	1.00	.00	3.00	3.00
5 (HW)	.10	.50	.32	.53	.00	.00	1.00	1.00
6 (HW)	1.20	.30	1.03	.48	.00	.00	3.00	1.00
7 (HW)	.80	.80	.92	.92	.00	.00	3.00	3.00
Average of Low Workload Missions	1.33	1.03	.58	.65	.75	.00	2.50	2.00
Average of High Workload Missions	.70	.53	.40	.32	.00	.00	1.33	1.00
Non-Repeated Query								
1 (LW)	1.10	1.22*	1.10	1.31	.00	.00	3.00	3.00
2 (LW)	1.50	1.88****	1.35	.83	.00	1.00	3.00	3.00
3 (LW)	1.80	1.61***	1.32	1.14	.00	.00	3.00	3.00
4 (LW)	1.40	1.55*	1.43	1.07	.00	.00	3.00	3.00
5 (HW)	2.07*	1.30	1.00	1.25	1.00	.00	3.00	3.00
6 (HW)	2.00	2.00*	1.05	1.15	1.00	.00	3.00	3.00
7 (HW)	1.70	1.97**	1.16	1.25	.00	.00	3.00	3.00
Average of Low Workload Missions	1.45	1.56	.59	.56	.75	.62	2.50	2.35
Average of High Workload Missions	1.92	1.76	.61	.71	1.00	.67	2.56	3.00

* Contained missing data for one team, which was replaced with the mission mean

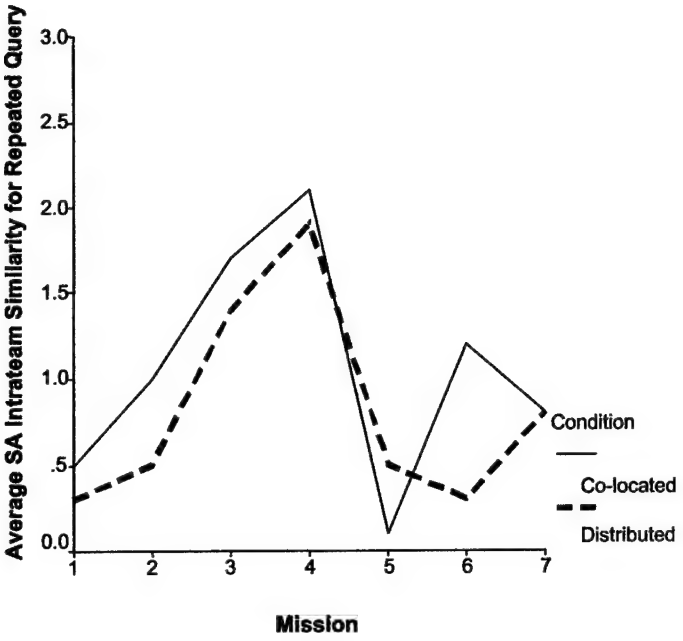


Figure 12. Average situation awareness intrateam similarity on the repeated query for both co-located and distributed teams at each mission.

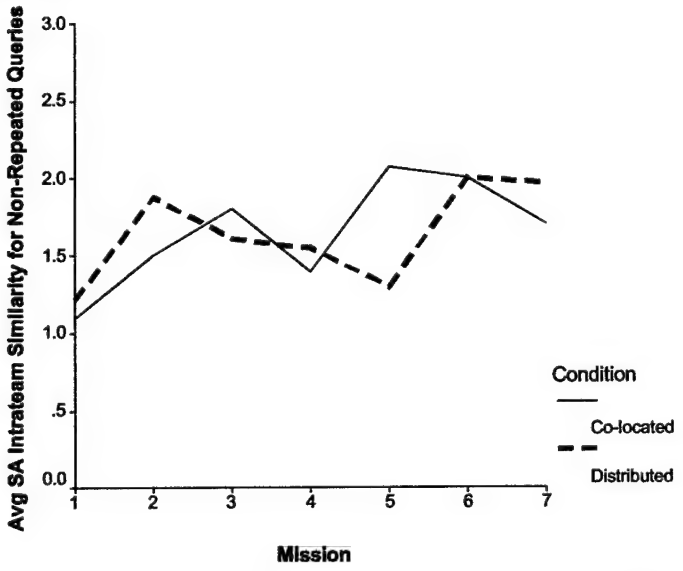


Figure 13. Average situation awareness intrateam similarity on the non-repeated queries for both co-located and distributed teams at each mission.

Post-hoc comparisons were conducted to examine the interaction between mission and query type. As Table 15 shows, intrateam similarity was significantly higher for the non-repeated queries than for the repeated query during Missions 1-2 and Missions 5-7.

Table 15
T Statistics for the Comparison of the Average Intrateam Similarity on the Repeated Query Minus Average Intrateam Similarity on the Non-repeated Queries at each Mission

Mission	*t* statistic	*p*-value
1	-2.64	.01
2	-2.92	.01
3	-.42	.68
4	1.41	.17
5	-4.90	.00
6	-3.97	.00
7	-3.12	.00

$df = 38$

Two planned contrasts were conducted to further analyze intrateam similarity. The contrasts were aimed at answering (1) whether teams' intrateam similarity improved over the low workload mission, and (2) whether there was an effect of workload on intrateam similarity. The effect of condition was omitted from the following contrasts since no significant effects of condition were found in the omnibus test. A univariate, repeated measures analysis of variance with two repeated factors (mission and query type) was used to analyze each contrast.

First, did intrateam similarity improve over the low workload missions? A comparison of Mission 1 and Mission 4 was performed in order to answer this question. Team responses to the situation awareness queries were significantly more similar in Mission 4 than Mission 1, $F(1, 19) = 20.32, p < .01$. Intrateam similarity did not differ as a function of query type, $F(1, 19) < 1$ but an interaction between mission and query type did emerge, $F(1, 19) = 10.56, p < .01$. *Post-hoc* comparisons of Mission 1 to Mission 4 revealed that for the repeated query, intrateam similarity significantly improved, $F(1, 19) = 51.75, p < .01$, but for the non-repeated query, intrateam similarity did not differ significantly from Mission 1 to Mission 4, $F(1, 19) < 1$.

For the second planned contrast, was intrateam similarity affected by workload? In a comparison of Mission 4 to Mission 5, an effect of workload was found, $F(1, 19) = 9.60, p < .01$, where intrateam similarity was lower for the high workload mission than for the low workload mission. Intrateam similarity was also lower for the repeated query than for the non-repeated queries, $F(1, 19) = 8.00, p = .01$. Furthermore, there was an interaction between workload and query type, $F(1, 19) = 12.97, p < .01$. *Post-hoc* comparisons of Mission 4 to Mission 5 showed that for the repeated query, teams were less similar in their responses during the high workload mission than the low workload mission, $F(1, 19) = 38.94, p < .01$. For the non-repeated query, there was no difference in intrateam similarity between low workload and high workload, $F(1, 19) < 1$.

Holistic situation awareness. Table 16 shows holistic situation awareness accuracy for co-located and distributed teams on a mission-by-mission basis. The table also shows an average of holistic accuracy over the low workload missions. As with accuracy and intrateam similarity, missing data (3 points) were replaced with the corresponding mission mean.

Table 16

Holistic Situation Awareness Accuracy on the Repeated Query and Non-Repeated Queries for Co-located and Distributed Teams

Mission	Mean		St. Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Repeated Query								
1 (LW)	.40	.10	.52	.32	.00	.00	1.00	1.00
2 (LW)	.20	.40	.42	.52	.00	.00	1.00	1.00
3 (LW)	.50	.50	.53	.53	.00	.00	1.00	1.00
4 (LW)	.77*	.57*	.42	.50	.00	.00	1.00	1.00
5 (HW)	.00	.10	.00	.32	.00	.00	.00	1.00
6 (HW)	.10	.00	.32	.00	.00	.00	1.00	.00
7 (HW)	.60	.10	.52	.32	.00	.00	1.00	1.00
Average of Low Workload Missions	.47	.39	.30	.28	.00	.00	1.00	.75
Average of High Workload Missions	.23	.07	.16	.14	.00	.00	.33	.33
Non-Repeated Query								
1 (LW)	1.00*	.90	.02	.32	.95	.00	1.00	1.00
2 (LW)	.70	.80	.48	.42	.00	.00	1.00	1.00
3 (LW)	.80	.90	.42	.32	.00	.00	1.00	1.00
4 (LW)	1.00	.90	.00	.32	1.00	.00	1.00	1.00
5 (HW)	1.00	1.00	.00	.00	1.00	1.00	1.00	1.00
6 (HW)	.70	.90	.48	.32	.00	.00	1.00	1.00
7 (HW)	.90	.90	.32	.32	.00	.00	1.00	1.00
Average of Low Workload Missions	.87	.88	.21	.18	.50	.50	1.00	1.00
Average of High Workload Missions	.87	.93	.17	.14	.67	.67	1.00	1.00

* Contained missing data for one team, which was replaced with the mission mean

A univariate, repeated measures analysis of variance with two repeated factors (mission and query type) and one between-subjects factor (condition) was used to analyze situation awareness holistic accuracy. Results from the omnibus test are presented first, followed by the results of two planned contrasts. The co-located/distributed manipulation did not produce a significant effect on holistic accuracy, $F(1, 18) < 1$. In contrast, a main effect of mission was found, $F(6, 108) = 3.87$, $p < .01$, indicating that holistic accuracy changed significantly over missions. A main effect of query type was also present, $F(1, 18) = 145.80$, $p < .01$. A comparison of Figures 14 and 15 illustrates how teams were more accurate in their holistic responses to the non-repeated queries than to the repeated query. Furthermore, a significant interaction between mission and query type was revealed, $F(6, 108) = 5.25$, $p < .01$. The interaction between condition and query type was not significant, $F(1, 18) = 2.27$, nor was the interaction between condition and mission, $F(6, 108) = 1.49$. Finally, the three-way interaction among condition, mission, and query type was also not significant, $F(6, 108) = 1.08$.

Post hoc comparisons were conducted in order to locate the source of the mission by query type interaction. Recall that with accuracy and intrateam similarity, scores on the non-repeated queries were only higher than scores on the repeated query for Missions 1-2 and Missions 5-7. However, as Table 17 shows, holistic accuracy on the non-repeated queries was significantly higher than holistic accuracy on the repeated query at every mission.

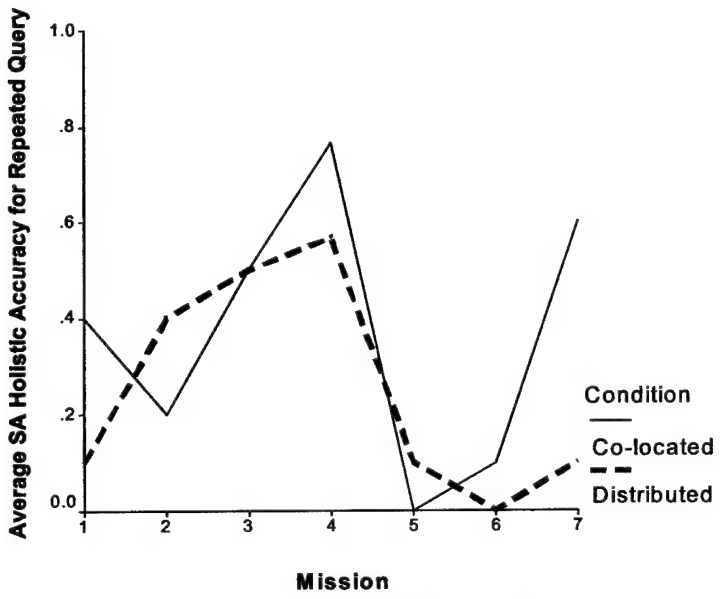


Figure 14. Average situation awareness holistic accuracy on the repeated query for both co-located and distributed teams at each mission.

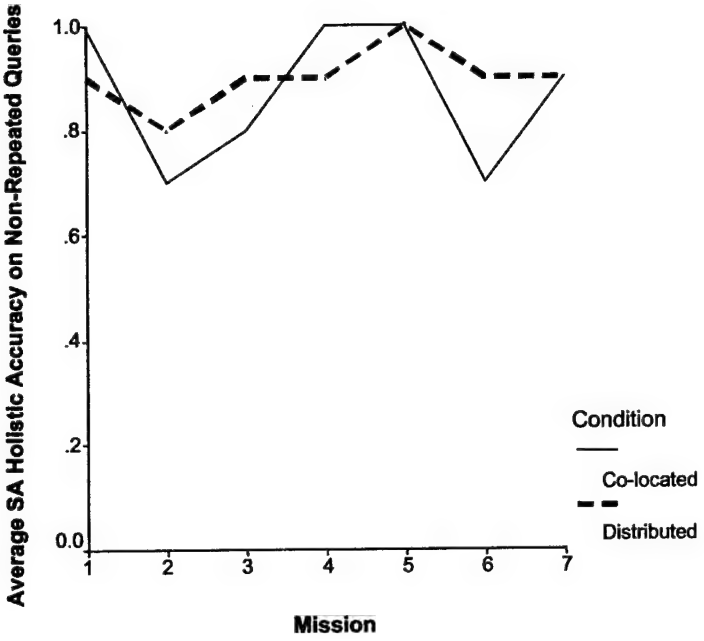


Figure 15. Average situation awareness holistic accuracy on the non-repeated queries for both co-located and distributed teams at each mission.

Table 17
Statistics for the Comparison of the Average Holistic Accuracy on the Repeated Query to the Average Holistic Accuracy on the Non-Repeated Queries at each Mission

Mission	*t* statistic	*p*-value
1	-6.27	.00
2	-3.11	.00
3	-2.48	.02
4	-2.48	.02
5	-19.00	.00
6	-7.18	.00
7	-4.26	.00

df = 38

Planned contrasts were conducted in order to answer the following questions: (1) Did teams' holistic accuracy improve over the low workload missions, and (2) was there an effect of workload on holistic accuracy? The effect of condition was omitted from these contrasts on holistic accuracy for the same reason it was excluded from the comparisons conducted on accuracy and intrateam similarity. That is, the lack of significant effects in the omnibus test justified omitting condition from further analyses. A univariate, repeated measures analysis of variance with two repeated factors (mission and query type) was used to analyze each contrast.

To determine whether holistic accuracy improved over the low workload missions, the scores from Mission 1 were compared to the scores from Mission 4. Teams' holistic accuracy scores improved significantly from Mission 1 to Mission 4, $F(1, 19) = 6.44$, $p = .02$. Teams' holistic

accuracy scores were also significantly higher for the non-repeated queries than for the repeated query, $F(1, 19) = 39.23, p < .01$. Finally, there was an interaction between mission and query type, $F(1, 19) = 10.10, p < .01$, (see Figure 15). *Post-hoc* comparisons showed that for the repeated query teams became significantly more accurate in their holistic responses by Mission 4, $F(1, 19) = 10.29, p < .01$. However, the teams' holistic responses to the non-repeated queries did not significantly change from Mission 1 to Mission 4, $F(1, 19) < 1$.

The final planned contrast examined the effect of workload on holistic accuracy by comparing the holistic accuracy during the final low workload mission (Mission 4) to holistic accuracy during first high workload mission (Missions 5). An effect of workload was found, $F(1, 19) = 18.89, p < .01$, where teams were significantly less accurate in reaching consensus to the situation awareness queries during the high workload mission. Teams were also significantly less accurate in reaching a consensus on the repeated query than the non-repeated queries, $F(1, 19) = 112.57, p < .01$. An interaction also emerged between workload and query type, $F(1, 19) = 42.26, p < .01$. *Post hoc* comparisons showed that holistic accuracy on the repeated query significantly declined during the high workload mission, $F(1, 19) = 33.77, p < .01$, but for the non-repeated query, holistic accuracy did not change significantly across the levels of workload, $F(1, 19) = 1.00$.

Correlations between Objective and Subjective Measures of Situation Awareness. In order to compare objective and subjective measures of situation awareness, the items on the SART questionnaire (see Appendix L), which specifically asked participants to rate their perception of how aware they were of the situation were correlated with situation awareness accuracy and holistic accuracy scores. Individual responses to the SART questionnaire were averaged across items and team members to estimate the teams' perception of their situation awareness. Average SART ratings from Mission 4 were used to estimate subjective situation awareness for low workload and ratings from Mission 5 were used to estimate ratings during the high workload missions.

Table 18 presents correlations between subjective and objective situation awareness scores (repeated and non-repeated queries) during low and high workload. Subjective situation awareness ratings taken at Mission 4 were significantly correlated with situation awareness accuracy on the non-repeated queries in high workload, indicating that teams who believed they had good situation awareness at Mission 4, were also more accurate on the non-repeated queries at Mission 5. Furthermore, subjective situation awareness ratings taken at Mission 5 were significantly related to (1) accuracy on the non-repeated queries in low workload, and (2) holistic accuracy on the non-repeated queries during high workload. In general, these relationships suggest that teams who reported having more situation awareness were also more accurate in their individual and team responses to the non-repeated queries.

Table 18
Correlations Between Subjective Situation Awareness Ratings and Situation Awareness Accuracy and Holistic Accuracy

	Repeated Accuracy (M4)	Repeated Accuracy (M5)	Repeated Holistic (M4)	Repeated Holistic (M5)	Non-Repeated Accuracy (LW)	Non-Repeated Accuracy (HW)	Non-Repeated Holistic (LW)	Non-Repeated Holistic (HW)
SART (M4)	.27	-.09	.08	-.36	.04	.43*	.26	.37
SART (M5)	-.09	-.04	-.09	.07	.51**	.09	.10	.44**

N = 20 * $p < .10$ ** $p < .05$

This rather weak pattern of correlations between objective and subjective measures of situation awareness may reflect on the subjective measure, which is often criticized on the grounds of being subjective. However, we also suspect that our objective situation awareness measure may reflect something other than situation awareness, especially at the team level. In particular, most queries involved information that was available to only one team member. Also, good teams could improve on the repeated query because it recurred and the team became better at estimating their own performance (i.e., number of targets that will be photographed in a mission). These issues come up again in the measurement evaluation section of the report.

To summarize:

- For all of the situation awareness measures (accuracy, intrateam similarity, and holistic accuracy), there was no effect of dispersion condition.
- Accuracy, similarity, and holistic accuracy improved between Missions 1 and 4 for repeated queries, but not for nonrepeated queries. Accuracy and holistic accuracy declined between Missions 4 and 5 for repeated queries, but not for nonrepeated queries

4.3.4 Taskwork Knowledge

The means and standard deviations as well as the minimum and maximum scores for *overall taskwork accuracy* during Knowledge Sessions 1 and 2 can be seen in Table 19 for distributed and co-located teams. Taskwork data collected during Knowledge Session 2 was missing for one team (Team 7). The means reveal that distributed teams did slightly better during Knowledge Session 1, whereas co-located teams did slightly better during Knowledge Session 2.

A mixed two-factor ANOVA revealed a significant interaction between condition and knowledge session, $F(1, 17) = 5.17, p = .04$. A main effect of knowledge session was also found, $F(1, 17) = 4.05, p = .06$, where overall accuracy was higher in Knowledge Session 2. There was no main effect of condition $F(1, 17) < 1$. As *post hoc* tests reveal, co-located teams improved in overall accuracy from Knowledge Session 1 to Knowledge Session 2, $F(1, 8) = 6.62, p = .03$, but distributed teams' overall accuracy scores did not change, $F(1, 9) < 1$. Additionally, there were

no differences between co-located and distributed teams at Knowledge Session 1, $F(1, 18) = 2.60$, or at Knowledge Session 2, $F(1, 17) = 1.39$.

Table 19
Overall Taskwork Accuracy in Co-located and Distributed Conditions for Knowledge Session 1 and Knowledge Session 2

Knowledge Session	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	.44	.48	.06	.04	.37	.41	.56	.53
2	.50	.47	.05	.04	.39	.40	.59	.54

Table 20 displays the descriptive statistics for *taskwork positional knowledge*. A mixed two-factor ANOVA revealed no significant interaction between condition and knowledge session, $F(1, 17) < 1$ nor a significant effect of condition, $F(1, 17) < 1$. There was also no significant difference across knowledge sessions in positional knowledge, $F(1, 17) = 1.97$.

Table 20
Taskwork Positional Knowledge in Co-located and Distributed Conditions for Knowledge Session 1 and Knowledge Session 2

Knowledge Session	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	-.19	-.07	.55	.60	-.96	-.96	.52	.57
2	.15	.13	.56	.65	-.51	-.94	1.18	1.15

Taskwork interpositional knowledge was also analyzed for both sessions as a function of the co-located/distributed manipulation. As with overall accuracy, there was a significant interaction between knowledge session and condition, $F(1, 17) = 3.29$, $p = .09$, as well as a significant main effect of knowledge session, $F(1, 17) = 6.09$, $p = .03$. No significant effect of condition was found, $F(1, 17) < 1$. Again, *post-hoc* tests confirmed that co-located teams drastically improved in interpositional knowledge across knowledge sessions, $F(1, 8) = 8.86$, $p = .02$, while distributed teams' interpositional knowledge did not significantly improve from Knowledge Session 1 to Knowledge Session 2, $F(1, 9) < 1$. However, there were no significant differences in interpositional knowledge between co-located teams and distributed teams at Knowledge Session 1, $F(1, 18) < 1$ or at Knowledge Session 2, $F(1, 17) = 2.40$. Across all conditions, teams achieved higher interpositional knowledge scores in Knowledge Session 2.

Table 21
Taskwork Interpositional Knowledge in Co-located and Distributed Conditions for Knowledge Session 1 and Knowledge Session 2

Knowledge Session	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	-.20	-.08	.55	.40	-.70	-.84	1.24	.62
2	.32	.00	.46	.43	-.62	-.90	.68	.49

We also tested *taskwork intrateam similarity*, for which the descriptive data are displayed in Table 22. There was no significant interaction between condition and knowledge session, $F(1, 17) = 2.67$, but a significant effect for session was revealed, $F(1, 17) = 14.39$, $p < .01$ with both co-located and distributed teams becoming more similar over time. There was no significant condition effect, $F(1, 17) < 1$.

Table 22
Taskwork Intrateam Similarity in Co-located and Distributed Conditions for Knowledge Session 1 and Knowledge Session 2

Knowledge Session	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	.36	.38	.06	.06	.30	.28	.49	.47
2	.43	.41	.07	.07	.34	.27	.56	.53

The final taskwork variable we examined was *holistic taskwork accuracy*. Descriptive data are displayed in Table 23. For this variable, there was a significant interaction between condition and session, $F(1, 16) = 12.27$, $p < .01$. A significant effect of session also emerged, $F(1, 16) = 3.07$, $p = .10$, indicating that across teams, holistic accuracy was higher at knowledge Session 2. There was no significant effect of condition, $F(1, 16) < 1$. *Post hoc* tests indicated that co-located teams became more accurate from Knowledge Session 1 to Knowledge Session 2 on the holistic measure, $F(1, 8) = -17.99$, $p < .01$, while distributed teams' holistic accuracy did not significantly change across sessions, $F(1, 8) = 1.24$. Furthermore, at Knowledge Session 1, distributed teams were significantly more accurate on the holistic ratings than co-located teams, $F(1, 18) = 4.39$, $p = .05$, whereas there was no significant difference in holistic accuracy between co-located and distributed teams at Knowledge Session 2, $F(1, 16) = 2.33$.

Table 23
Holistic Taskwork Accuracy in Co-located and Distributed Conditions for Knowledge Session 1 and Knowledge Session 2

Knowledge Session	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	.53	.59	.07	.05	.39	.50	.63	.69
2	.62	.56	.06	.08	.52	.44	.71	.71

To summarize:

- With the exception of a single specific effect (i.e., distributed teams having higher holistic taskwork scores than co-located teams at Session 1) there were no differences in taskwork knowledge of teams due to dispersion.

- With the exception of positional knowledge, there was general improvement in taskwork knowledge scores from Session 1 to 2. This improvement is mostly attributable to co-located teams (however both co-located and distributed teams became more similar over sessions).

4.3.5 Teamwork Knowledge

The means and standard deviations as well as the minimum and maximum scores for *teamwork overall accuracy* during Knowledge Session 1 and Knowledge Session 2 are given in Table 24 for distributed and co-located teams. The means reveal that both co-located and distributed teams scored higher on overall teamwork knowledge in Session 2 than in Session 1. A mixed two-factor ANOVA revealed no significant interaction between condition and knowledge session, $F(1, 18) < 1$ nor a significant effect of condition, $F(1, 18) < 1$. There was a significant increase across knowledge sessions, $F(1, 18) = 8.44$, $p = .01$, with both co-located and distributed teams obtaining higher teamwork knowledge scores in Session 2.

Table 24
Teamwork Overall Accuracy in Co-located and Distributed Conditions for Knowledge Session 1 and Knowledge Session 2

Knowledge Session	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	22.90	23.07	2.47	2.41	17.33	18.00	25.33	27.33
2	24.87	24.77	1.43	1.57	23.33	22.67	28.00	27.67

Knowledge of one's own role or *positional knowledge* (AVO, PLO, or DEMPC) as well as knowledge of other roles (*inter-positional knowledge*) were also examined for teamwork. Descriptive statistics for these variables are displayed in Tables 25 and 26.

Table 25
Teamwork Positional Knowledge in Co-located and Distributed Conditions for Knowledge Session 1 and Knowledge Session 2

Knowledge Session	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	.72	.72	.09	.11	.55	.47	.81	.88
2	.77	.79	.09	.05	.62	.71	.87	.89

Table 26
Teamwork Inter-Positional Knowledge in Co-located and Distributed Conditions for Knowledge Session 1 and Knowledge Session 2

Knowledge Session	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	.70	.71	.07	.08	.61	.59	.82	.81
2	.77	.74	.07	.09	.61	.61	.87	.88

Values are based on percentage correct because the number of items on which a score was based varied by role. A mixed two-factor ANOVA revealed significant effects of knowledge session on both positional, $F(1, 18) = 6.53$, $p = .02$ and inter-positional knowledge, $F(1, 18) = 3.35$, $p = .08$, with teams apparently having more knowledge in Session 2 than in Session 1. However, there

was no interaction with condition for positional, $F(1, 18) < 1$, or inter-positional knowledge, $F(1, 18) < 1$, nor was there a main effect of condition for positional, $F(1, 18) < 1$, or inter-positional knowledge, $F(1, 18) < 1$.

As can be seen in Table 27, *intrateam similarity* also improved from Knowledge Session 1 to Knowledge Session 2 for both co-located and distributed teams. Thus teams achieved greater similarity in Session 2 than in Session 1, $F(1, 18) = 53.37$, $p < .01$. However, as with teamwork knowledge, there was no significant interaction between session and condition, $F(1, 18) = 2.45$, $p = .14$, nor was there a significant effect of condition, $F(1, 18) < 1$.

Table 27
Teamwork Similarity in Co-located and Distributed Conditions for Knowledge Session 1 and Knowledge Session 2

Knowledge Session	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	8.30	6.80	2.00	2.62	6.00	2.00	12.00	12.00
2	11.60	11.90	2.01	2.69	9.00	6.00	14.00	15.00

Holistic teamwork accuracy means in Table 28 show that co-located teams had poorer holistic teamwork knowledge in Session 2 than in Session 1, whereas distributed teams had more knowledge in Session 2 than in Session 1. However, there was no significant interaction between condition and session, $F(1, 18) = 2.77$, nor was there a significant effect of condition, $F(1, 18) < 1$. Furthermore, there was no significant improvement in holistic teamwork knowledge between Sessions 1 and 2, $F(1, 18) < 1$.

Table 28
Holistic Teamwork Accuracy in Co-located and Distributed Conditions for Knowledge Session 1 and Knowledge Session 2

Knowledge Session	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	26.60	25.60	1.65	2.41	24	23	28	30
2	25.80	26.40	2.35	1.84	23	22	29	28

To summarize:

- There were no significant differences between the co-located and distributed teams on the teamwork knowledge measures.
- Both co-located and distributed teams improved between Knowledge Session 1 and Knowledge Session 2 on overall accuracy, positional knowledge, and interpositional knowledge, but did not improve on the holistic knowledge measure. Team members also became more similar in their ratings over the two sessions.

4.3.6 Correlations of Performance and Process

We report correlations separately for co-located and distributed teams for *critical incident process* (Table 29) because we found significant differences between co-located and distributed teams using the critical incident measure. Correlations involving the summary process measure (Table 30) are also broken down by dispersion condition for Mission 5 because there were significant differences between co-located and distributed teams on the performance measure for this mission. Missions 4 and 5 are used to represent the low workload and high workload missions respectively because team performance and process significantly varied by mission in the low workload missions.

Table 29
Correlations Between Team Performance and Critical Incident Process Scores for Co-located and Distributed Teams

Mission	Co-located	Distributed
4	.03	.88*
5	.05	.27

df = 8, * $p < .10$

Table 30
Correlations Between Team Performance and Summary Process Scores

Mission	All Teams	
4	.06	
5	Co-located	Distributed
	.36	.59*

df = 8, * $p < .10$

To summarize: Distributed teams with better process scores obtained higher scores on the UAV task in Mission 5. Distributed teams that had higher critical incident scores also performed better in the last low workload mission (Mission 4).

4.3.7 Correlations between Knowledge Measures and Performance or Process

In Experiment 1 there were 16 separate knowledge measures considered. In this analysis, taskwork and teamwork knowledge measures were considered at Knowledge Session 2 only. Each of these measures was scored against overall, positional, and interpositional referents as well as for similarity, yielding a total of eight taskwork and teamwork measures. Situation awareness involved a total of eight measures with four each for repeated and nonrepeated queries. The four included situation awareness accuracy and similarity each scored in low (Mission 4) and high (Mission 5) workload missions.

In order to summarize the correlations among the 16 knowledge measure variables, they were subjected to a hierarchical cluster analysis utilizing the centroid linkage method. Using Pearson correlations significant at $p \leq .10$ as a cluster cutoff, twelve variables formed six distinct, non-overlapping clusters. The remaining four factors did not enter into a cluster. Table 31 presents the clusters and the knowledge measures that form them.

Table 31
Clusters Among Knowledge Measures for Experiment 1

Cluster Name	Variables
1) Taskwork Accuracy-IPK	Taskwork Overall Accuracy Taskwork Interpositional Knowledge
2) Taskwork Role-Similarity	Taskwork Similarity Taskwork Positional Knowledge
3) Teamwork	Teamwork Accuracy Teamwork Positional Knowledge
4) Teamwork IPK-SA	Teamwork Interpositional Knowledge SA Similarity Repeated High Workload
5) SA Non-Repeated Low Workload	SA Accuracy Non-Repeated Low Workload SA Similarity Non-Repeated Low Workload
6) SA Non-Repeated High Workload	SA Accuracy Non-Repeated High Workload SA Similarity Non-Repeated High Workload

Relationship among knowledge clusters and team performance. To correlate each cluster with team performance, the variables within each cluster were standardized (if not already scaled) and averaged. Correlations between the clusters and team performance as well as between the four single variables and team performance can be seen in Table 32. A moderately significant correlation between teamwork IPK-SA (cluster 4) and performance indicated that teams with more interpositional role knowledge and situation awareness performed better in high workload than teams with lower levels of interpositional role knowledge and situation awareness. A second moderately significant correlation involving the single variable situation awareness similarity indicated that teams with more similar responses to situation awareness repeated queries performed better in low workload missions.

Table 32
Correlations Between Knowledge Measures Clusters and Team Performance

Cluster/Variable	Low Workload Performance	High Workload Performance
Cluster 1 - Taskwork Accuracy-IPK	.04	-.03
Cluster 2 - Taskwork Role-Similarity	-.04	.03
Cluster 3 - Teamwork	-.17	-.09
Cluster 4 - Teamwork IPK-SA	.03	.38*
Cluster 5 - SA Non-Repeated Low Workload	.12	.27
Cluster 6 - SA Non-Repeated High Workload	.22	.12
Teamwork Similarity	-.18	.24
SA Accuracy Repeated High Workload	-.20	-.06
SA Accuracy Repeated Low Workload	.35	-.04
SA Similarity Repeated Low Workload	.38*	-.16

* $p = .10$ ** $p < .01$ $df = 18$

Relationship between knowledge clusters and team process. As mentioned above, knowledge variables within each cluster were standardized and averaged in order to correlate each cluster with process. Correlations between the knowledge measures (clusters and single variables) and critical incident process are presented in Table 33. The taskwork role-similarity cluster (cluster 2) was found to be positively associated with critical incident process during high workload for distributed teams. This suggests that distributed teams with good process behaviors at critical times during the missions tended to exhibit higher levels of taskwork role knowledge and taskwork knowledge similarity at Knowledge Session 2. Another significant correlation emerged between teamwork (cluster 3) and distributed teams' critical incident process score during high workload which indicates that distributed teams that exhibited poor process behaviors at critical times during the missions tended to exhibit good teamwork accuracy and role knowledge at Knowledge Session 2.

Table 33
Correlations Between Knowledge Measures Clusters and Critical Incident Process

Cluster/Variable	Low Workload CIP	High Workload CIP	
		Co-located	Distributed
Cluster 1 - Taskwork Accuracy-IPK	.16	-.1	.26
Cluster 2 - Taskwork Role-Similarity	-.05	.03	.65*
Cluster 3 - Teamwork	-.08	.13	-.70*
Cluster 4 - Teamwork IPK-SA	-.14	.43	-.49
Cluster 5 - SA Non-Repeated Low Workload	.30	.38	.18
Cluster 6 - SA Non-Repeated High Workload	.06	-.22	.29
Teamwork Similarity	-.19	.15	-.42
SA Accuracy Repeated High Workload	.03	.20	-.22
SA Accuracy Repeated Low Workload	-.03	-.42	.33
SA Similarity Repeated Low Workload	.14	-.43	.17

*$p \leq .05$ $df = 18$ (low workload), $df = 8$ (high workload)

Correlations between the knowledge clusters and summary process can be seen in Table 34. The taskwork role-similarity cluster (cluster 2) was moderately correlated with summary process in the low workload missions indicating that teams with higher summary process scores exhibited poorer taskwork role knowledge and taskwork similarity at Knowledge Session 2. A highly significant correlation occurred between teamwork (cluster 3) and summary process indicating that teams in the low workload condition that demonstrated good summary process also exhibited good teamwork accuracy and role knowledge at Knowledge Session 2. Significant correlations were found between Teamwork IPK-SA (cluster 4) and summary process indicating that teams that had good situation awareness and for the repeated queries and interpositional teamwork knowledge also had good process in both low and high workload. Lastly, a moderately significant correlation was found between SA accuracy for repeated high workload queries and summary process indicating that teams that had good situation awareness in high workload missions exhibited poor summary process scores in these missions.

Table 34
Correlations Between Knowledge Measures Clusters and Summary Process

Cluster/Variable	Low Workload Summary Process	High Workload Summary Process
Cluster 1 - Taskwork Accuracy-IPK	.22	.20
Cluster 2 - Taskwork Role-Similarity	-.39*	.18
Cluster 3 - Teamwork	.50**	.06
Cluster 4 - Teamwork IPK-SA	.46**	.41*
Cluster 5 - SA Non-Repeated Low Workload	.15	.15
Cluster 6 - SA Non-Repeated High Workload	.03	.03
Teamwork Similarity	.29	.23
SA Accuracy Repeated High Workload	-.21	-.44*
SA Accuracy Repeated Low Workload	-.21	-.11
SA Similarity Repeated Low Workload	-.11	-.09

*$p \leq .10$ **$p < .05$ $df = 18$

4.4 Experiment 1: Discussion

In this experiment the effect of co-located versus distributed mission environments on team performance, process, and cognition was investigated. The team task was a UAV reconnaissance task and involved three individuals who worked together in seven 40-minute missions, the last three under higher workload than the first four. Each main dependent measure was analyzed in order to address the four hypotheses previously raised. The results are summarized in Table 35 in terms of answers to three main questions: (1) Was dispersion detrimental, (2) Was there early improvement (i.e., learning), and (3) Was increased workload detrimental.

Table 35
Summary of Experiment 1 Results

Measure	Was dispersion detrimental?	Was there early improvement (i.e., learning)?	Was increased workload detrimental?
Team performance	No, but slight benefit of distributed in later missions	Yes	Yes
Team Process	Yes, co-located had better CIP	Yes	Yes for CIP, no for SUM
Situation Awareness	No	Yes, repeated queries only	No
Taskwork Knowledge	No	Yes, but mostly for co-located	N/A
Teamwork Knowledge	No	Yes	N/A

Unexpectedly, geographic dispersion was not detrimental to team performance in our synthetic task environment. Nor was it detrimental to learning or in high workload environments. In fact, distributed teams performed slightly better than co-located teams under conditions of high workload. Thus, the first hypothesis (H1.1) regarding poorer performance for distributed teams was not supported.

We also hypothesized that there would be process deficits associated with the distributed environment during task acquisition that would drive early performance deficits as well as knowledge and situation awareness deficits. This hypothesis was partially supported by the critical incident process measure, which indicated superior process for co-located teams compared to distributed teams. This difference occurred for the acquisition period as well as for high workload missions. Further analysis of the items revealed that the primary difference between groups was in planning and adaptive behaviors, which the co-located teams seemed to carry out naturally. The distributed teams did not communicate as much in general and tended not to debrief after a mission. However, despite these process differences, there were generally no dispersion effects for the taskwork knowledge metrics, teamwork knowledge, or situation awareness measures resulting in only partial support for our hypothesis about process and knowledge deficits associated with dispersion (H1.2).

On the other hand, our teamwork and taskwork knowledge data seemed erratic with very few statistically significant knowledge-performance or knowledge-process correlations. We attribute this at least partially to the inappropriate placement of our knowledge sessions in this experiment. Session 1 was immediately following the Power Point and hands-on training, and before any mission experience. In retrospect, it may have been too early for participants to provide meaningful knowledge responses. The second knowledge session occurred after the seventh mission and immediately prior to the participants' departure. Although the timing of Session 2 was likely a better indicator of knowledge, it would have been better if it were the only session and placed somewhat earlier. We observed participants rushing through the knowledge items. In sum, this unfortunate placement of sessions could have resulted in excessively noisy data for the taskwork and teamwork measures in this experiment. Therefore the knowledge session in Experiment 2 was placed after the third of five missions.

Further, though increased workload did seem to have detrimental effects on performance, process, and situation awareness, these effects were not dependent on condition and therefore our third hypothesis (H1.3) was also not supported.

Our relatively low power, coupled with variance due to individual and team composition differences, may play a role in masking other interesting findings in this setting. Our fourth hypothesis (H1.4) concerned the contribution of these individual differences among teams as a moderator of process and performance effects. In fact, some of the contributions of individual differences within groups can be seen in the analysis of gender and working memory as factors in team composition. To illustrate we have rank ordered the Experiment 1 teams in terms of team performance averaged across the seven missions (See Table 36). Team 20 was excluded due to missing data so that there are only 19 teams. Note that co-located teams either perform very well or very poorly, while distributed teams tend to cluster in the center of the distribution.

Questions about the low-scoring co-located teams led us to explore some of the individual and team differences data more fully. It turns out that some variance in team performance is due to gender composition of teams with mixed-gender teams performing more poorly ($M = 345$) than same gender teams ($M = 390$). A Chi Square test of mixed vs. same gender by high vs. low scoring teams indicated that this difference is significant, $X^2(1, N=19) = 3.81, p = .05$.

In addition, working memory capacity of individuals, one of our secondary measures, seems to account for additional team performance variance as is reported in detail in a later section. The working memory task that was used in our study consisted of 32 items. Each item presented the participant with four to seven words and required them to remember the last three words in order. The working memory task yields component scores, one of which is the number of correct responses on items that require a mental transformation. On these items, the participant must remember the antonym of the word that is presented. For example, if the participant sees cold, the word hot should be stored and retrieved after the list of words has been presented. The working memory task yields a separate score for each member of the team and was administered on an individual basis before the team task began.

Table 36
Teams Ranked in Order (Lowest to Highest) on Team Performance

Team ID	Team Performance	C= Co-located; D=Distributed	Gender Composition	Team Working Memory Score (bold, italics = under median)
14	**294**	**COL**	**Mixed**	***50***
3	**305**	**COL**	**Mixed**	***50***
5	335	DIST	Mixed	59
13	**337**	**COL**	**Mixed**	***51***
8	339	**COL**	**Mixed**	61
17	341	DIST	Mixed	***42***
19	342	DIST	Mixed	67
6	343	DIST	Mixed	57
12	354	**COL**	**Mixed**	57
7	357	**COL**	Same	***41***
15	358	DIST	Mixed	59
21	362	DIST	Mixed	***55***
4	365	DIST	Mixed	***48***
16	370	**COL**	Same	***23***
9	375	DIST	Mixed	***53***
11	376	**COL**	**Mixed**	69
1	415	**COL**	Same	63
10	418	DIST	Same	60
2	428	**COL**	Same	62

If teams are examined on the basis of working memory scores (average among the three team members) and gender composition, we see that Teams 3, 13, and 14 are the only co-located teams that have both mixed gender composition and a low working memory team score (i.e., below a median cutoff; see Table 36). In other words, these co-located teams lacked both the gender composition and working memory capacity associated with high performing co-located teams. As an illustration of the importance of such differences, performance across all seven missions is plotted in Figure 16 for the distributed teams and these two groups of co-located teams. When these three teams are removed from the analysis, the co-located team performance mean across all missions is 377 compared to 359 for the distributed teams. Whereas this difference is not significant, the low workload team performance difference of 394 for remaining co-located teams and 346 for distributed teams is significant, $t(15) = 3.10, p < .01$. The difference for high workload missions (co-located $M = 353$, distributed $M = 376$) is not significant, and distributed teams still seem to hold the advantage.

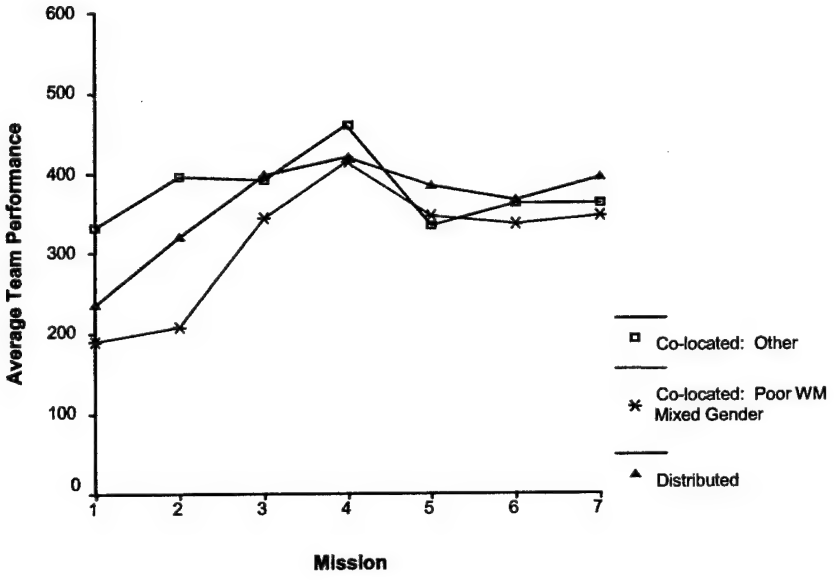


Figure 16. Team performance for distributed teams, three co-located teams (mixed gender and low working memory), and remaining co-located teams.

This analysis reveals the difference that team composition and individual differences can make in a set of findings. Perhaps, co-located teams would be superior to distributed teams if these differences were better controlled. Therefore in the next experiment we decided to focus on individual and team composition contributions to team performance and to hold gender composition constant by collecting data from only all-male teams.

4.5 Experiment 2: Team Cognition in Distributed Mission Environments

Experiment 2 was designed to address the first objective of this project which was to conduct empirical studies to investigate the impact of geographic dispersion and varying workload on team performance, process, and cognition in the context of the CERTT Lab's three-person UAV-

STE. In this section we describe Task 3 under the first objective of this project which is: Based on the data from the first study, design and collect data from a second experiment to investigate the combined effects of communication mode differences, familiarity, and co-presence on team cognition, process, and performance during task acquisition and skilled performance under varying workload conditions.

Results from Experiment 1 suggested that the distributed mission environment (as opposed to a co-located mission environment) had a negative impact on team process behavior. In particular, team process behavior involving mission planning and adaptive behavior was superior in co-located teams as indicated by our critical incident process measure. However, there was minimal impact on team performance or knowledge. There were some hints of effects on these variables, however, co-located teams, but not distributed, showed improvement on some of the taskwork knowledge measures over two sessions. Knowledge measures were also weakly, if at all, correlated with team performance and process, however there may have been problems with the placement of the knowledge sessions in the experimental protocol. Interestingly distributed teams tended to outperform co-located teams in later high workload missions. We also identified some outlying co-located groups with mixed gender and low working memory who tended to lower the mean performance of co-located teams on early trials. These hints of findings together with variance attributed to individual and team composition factors and problematic placement of knowledge sessions motivated Experiment 2. Experiment 2 was basically a replication of Experiment 1 using all male teams and a single knowledge session strategically placed after the third mission. In addition, number of missions was reduced from seven to five in order to make it possible to collect data in a single session. Some measures used in Experiment 1 were also dropped in Experiment 2 for similar reasons. In addition, not only was working memory measured, but verbal processing speed as well (these are discussed in a later archival analysis section).

The following hypotheses are based on the assumptions stated previously regarding factors associated with DMEs, as well as our theoretical views concerning the relations between team cognition, process, and performance.

H2.1 During task acquisition DME teams will suffer process deficits resulting in slower acquisition rates and overall poorer acquisition performance compared to teams in the co-located condition.

H2.2 During task acquisition DME teams will suffer process deficits resulting in slower development of team knowledge and situation models compared to teams in the co-located condition.

H2.3 Although by later trials, DME teams may "catch up" in terms of team cognition and performance to co-located teams, and may compensate for process deficits during low workload periods, process deficits, and consequently performance and situation model deficits, will occur in periods of high workload.

H2.4 Individual differences among DME teams in terms of process strategy may moderate any deleterious effects of the DME, such that the "best" DME teams can overcome DME limitations compared to DME teams with poorer team process.

4.6 Experiment 2: Method

4.6.1 Participants

Twenty three-person teams of NMSU students voluntarily participated in one seven-hour session. Individuals were compensated for their participation by payment of $6.00 per person hour with each of the three team-members on the highest-performing team receiving a $50.00 bonus. All of the participants were males. Most of the participants were either Caucasian (43%) or Hispanic (33%). Participants ranged in age from 17 to 42. The participants were randomly assigned to teams and to role (AVO, PLO, or DEMPC).

4.6.2 Equipment and Materials

The study took place in the CERTT Laboratory configured for the UAV-STE described previously. For the most part, materials were the same as those used in Experiment 1 with the exception of minor changes in measurement materials. Custom software was developed to gather the teamwork individual and consensus ratings electronically. SART, secondary knowledge questions, social desirability, and the emerging leadership survey were not used in Experiment 2. Verbal processing speed was added as a measure in Experiment 2. The experimenter control station was also upgraded by the addition of a fifth computer, which allowed experimenters to take control of any of the six participant computers. This enabled experimenters to start and terminate mission applications remotely from the experimenter room. Finally, the original Javelin Systems quad splitter, which allowed video input from each of the 4 cameras to be displayed simultaneously on the monitor, was upgraded with a Sensormatic Monochrome quad splitter.

4.6.3 Measures

Details of all of the measures used in Experiment 2 are described in the primary and secondary measures sections of Experiment 1. Performance, process, and knowledge measures (including situation awareness) were administered and scored identically to Experiment 1. Of the secondary measures used in Experiment 1, SART, secondary knowledge questions, social desirability, and the emerging leadership survey were not administered in Experiment 2. These measures had been administered in Experiment 1 for specific purposes (e.g., secondary knowledge questions were used to provide additional measures for the multi-trait multi-method analysis) or they were discontinued due to time constraints and lack of interesting results (e.g., social desirability, SART). A new measure of verbal processing speed (described in the Experiment 1 secondary measure section) was administered for the first time in Experiment 2.

4.6.4 Procedure

Experiment 2 consisted of one seven-hour session with five, instead of seven, 40-minute missions and one, instead of two, knowledge elicitation sessions (see Table 37). The post

mission questionnaire contained only the NASA TLX questions (not SART) and was only administered after Missions 4 and 5. The knowledge sessions proceeded as follows: taskwork ratings, taskwork consensus ratings, teamwork ratings, and teamwork consensus ratings with no secondary knowledge questions. Breaks were given before Mission 1, before Mission 3, and before Mission 4. Other than these changes, procedures were identical to those of Experiment 1.

Table 37
Experiment 2 Protocol

SESSION 1
Working Memory Measure
Task Training
Mission 1 (low workload)
Mission 2 (low workload)
Mission 3 (low workload)
Knowledge Measures
Mission 4 (low workload)
Post Mission Questionnaire
Mission 5 (high workload)
Post Mission Questionnaire
Debriefing Questions

4.7 Experiment 2: Results

As stated earlier, team performance, team process behaviors, and knowledge measures (including knowledge relevant to situation awareness) are the focus of this project and are reported in the results section that follows. Results are summarized at the end of each section to facilitate an understanding of the main points. Some detailed analyses of workload measures are presented in the appendix (see Appendix S).

4.7.1 Team Performance

Table 38 shows the means for the co-located and distributed teams for each mission and Figure 17 provides a graph of these means. Although distributed teams had slightly higher team performance scores than co-located teams on all but the last mission, a mixed ANOVA revealed no main effect of the co-located/distributed manipulation, $F(1, 18) < 1$, nor an interaction between the dispersion condition and mission, $F(4, 72) < 1$. However, there was a detectable effect of mission, $F(4, 72) = 33.47$, $p < .01$.

Table 38
Team Performance in Co-located and Distributed Conditions

Mission	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	282	289	85	62	174	182	414	373
2	362	372	75	87	220	223	476	459
3	409	439	83	77	287	343	532	573
4	419	437	106	89	273	322	580	581
5	359	335	77	69	268	228	461	430

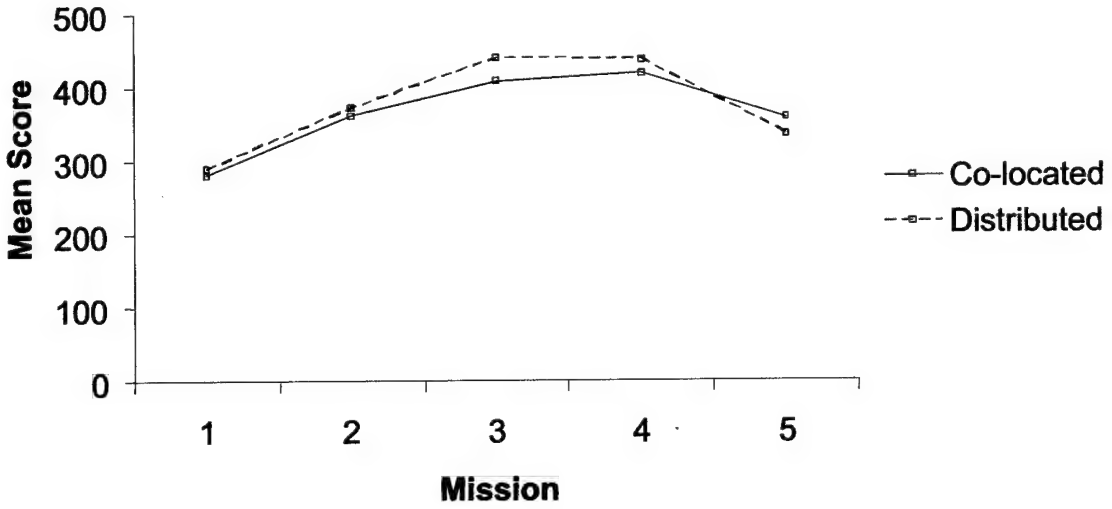


Figure 17. Performance scores for co-located and distributed teams

Sequential acquisition contrast effects are shown in Table 39. An analysis with a contrast between Missions 1 and 4, and its interaction with co-location, indicated that teams in both conditions learned the task. As with Experiment 1, a comparison of Missions 1 and 4 revealed that performance improved during the low workload missions, $F(1, 18) = 88.15$, $p < .01$, (see Table 38 for Means and SDs), with no interaction between missions and the dispersion condition $F(1, 18) < 1$.

Table 39
Sequential Acquisition Contrast Effects for Performance (Means are Adjusted for the Repeated Measures Model)

Contrast Between Missions	*B* (mean difference)	SE_B	β	*t*	*p*
2 - 1	84.51	9.11	0.57	9.28	<.01
3 - 2	88.05	11.15	0.59	7.89	<.01
4 - 3	34.39	11.15	0.23	3.08	<.01
5 - 4	-23.17	9.11	-0.16	-2.54	0.01

Comparisons were also made between performance in Mission 4 and the Mission 4 performance of teams from an earlier experiment (Cooke, et. al., 2001) to determine whether the teams reached asymptote in Mission 4 as they had in the earlier experiment. A two degree of freedom test including co-located teams against earlier teams, and distributed teams against earlier teams, produced no detectable difference, $F(2, 28) < 1$, in Mission 4 performance, so we assume that both co-located and distributed teams in the present experiment reached asymptote in Mission 4.

An increase in workload between Missions 4 and 5 appears to have produced a decline in performance (see Table 38 for Means and SDs). Teams (averaged across dispersion condition) performed better in the last low workload mission (Mission 4) than in the high workload mission (Mission 5), $F(1, 18) = 31.49, p < .01$, but there was no interaction between mission and dispersion condition, $F(1, 18) = 2.22$. We also compared co-located and distributed teams on the high workload mission (Mission 5). Unlike Experiment 1, there was no detectable difference between co-located and distributed teams on the high workload mission, $t(18) = .74$.

To summarize:

- Co-located and distributed teams learned the task during the low workload missions and performed more poorly when workload was increased.
- No effect of dispersion condition on team performance was found.

In general, the team performance results found in Experiment 1 were replicated in Experiment 2 with all-male teams and our hypothesis regarding performance deficits of DME teams was not supported by our findings.

4.7.2 Team Process

To calculate agreement between the two process raters, we computed a scaled proportion of agreement index (Po(scale); Cooke, et. al., 2001). Between all pairs of raters, we computed the absolute value of the deviation, scaled to the range of the possible scores. This normalized disagreement measure was then subtracted from 1, yielding:

Po(scale) = 1 - |Rater1 - Rater2| / Range.

Next, for each mission, we tested the Po(scale) for each process measure (i.e., critical incident process and summary process) using a one-sample t-test, against 0. Every process measure at every mission (both critical incident and rating measures) was detectably larger than disagreement, with no p-value being larger than .01, and almost all of them being smaller than .00 (see Appendix U). Therefore, agreement was adequate for the process measures, and we averaged between the two raters to yield an overall process score for each item.

Critical incident process. Prior to combining individual critical incident process items into an overall score, we tested whether individual items were additive in terms of an overall proportion. As we did for Experiment 1 critical incident process items, we performed a hierarchical centroid clustering on the item scores obtained in Experiment 2 using the 1 – correlation distance metric.

Figure 18 depicts the distances between centroids at each clustering stage. These distances tended to vary more for the Experiment 2 items than they did in Experiment 1. However the linear fit of the distance differences across clusters was judged to be sufficient in order to combine individual items into an overall proportion. As with Experiment 1, if a particular item was missing its value was subtracted from the proportion denominator. Table 40 gives the descriptive statistics of the overall critical incident score for co-located and distributed teams at each mission.

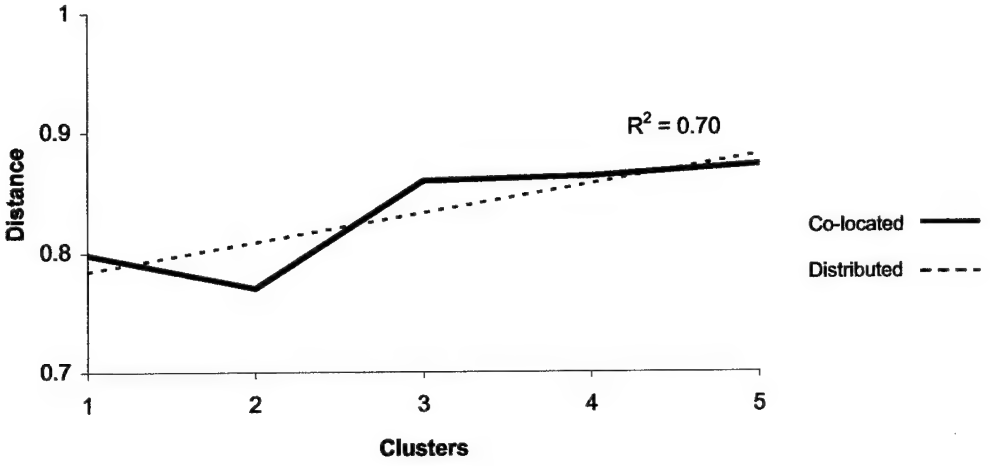


Figure 18. Experiment 2 critical incident process items; distance by cluster.

Table 40
Team Critical Incident Process Scores for Co-located and Distributed Conditions

Mission	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	.46	.42	.16	.14	.20	.19	.70	.64
2	.54	.57	.21	.16	.25	.35	1.00	.95
3	.61	.51	.13	.10	.40	.40	.85	.70
4	.59	.54	.17	.15	.30	.35	.95	.85
5	.51	.43	.20	.14	.10	.19	.75	.60

For Experiment 2 critical incident process we employed the same data analysis as for Experiment 1 with the number of missions decreasing from 7 to 5. Specifically we tested for main effects of mission and condition, as well as the interaction effect between the two. The design was a 2 X 5 (Condition X Mission) mixed design with condition as a between subjects factor and mission as a within subjects factor. There were 10 teams per condition, thus N = 2 X 10 X 5 = 100. The planned comparisons were identical to Experiment 1 analysis, specifically comparisons using only Missions 1 and 4 (acquisition tests) and comparisons using only Missions 4 and 5 (workload tests) were made.

The analysis identified a main effect of mission, $F(4, 72) = 4.29, p < .01$. This difference across missions can be seen in Figure 19. The main effect of condition however was not significant, $F(1, 18) < 1$. As depicted in Figure 19, although critical incident process scores tended to be higher for co-located than for distributed teams this difference was not statistically significant. In addition the interaction effect between condition and mission was not significant, $F(4, 72) < 1$. Apparently, any differences in critical incident process scores across missions did not depend on the particular dispersion condition.

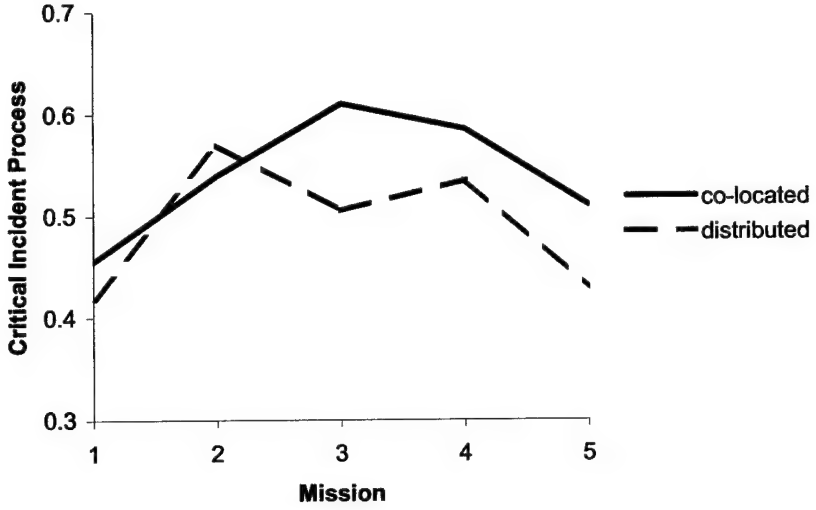


Figure 19. Mean Experiment 2 co-located and distributed critical incident process scores over missions.

The planned comparisons for acquisition (between Missions 1 and 4) revealed a significant main effect of mission, $F(1, 18) = 15.91, p < .01$. Therefore we conclude that teams in both conditions had significantly higher critical incident process scores in Mission 4 relative to Mission 1 (see Figure 19). The acquisition main effect of condition was not significant, $F(1, 18) < 1$. Thus, over the two acquisition mission levels, critical incident process scores were not statistically different between co-located and distributed teams. The acquisition interaction effect was also insignificant, $F(1, 18) < 1$, indicating that the observed statistical mission effect did not depend on condition. From these comparisons we conclude that both co-located and distributed Experiment 2 teams show acquisition of good critical incident process and that these acquisition curves were not statistically different.

The workload planned comparisons (between Missions 4 and 5) also found a significant main effect of mission, $F(1, 18) = 4.74, p < .05$. As can be seen in Figure 19, both co-located and distributed teams show decreased critical incident process in Mission 5 (high workload) relative to Mission 4 (low workload). The workload main effect of condition was not significant, $F(1, 18) = 1.18$. Thus critical incident process averaged over Missions 4 and 5 were not statistically different between co-located and distributed teams. Likewise the workload interaction effect was not significant, $F(1, 18) < 1$. Apparently the Mission 5 critical incident drop off was similar for both co-located and distributed teams.

From the results of the planned comparisons we reached some conclusions that seem to point to a consistent finding between Experiments 1 and 2. Regardless of condition, teams tend to improve on their critical incident process behaviors between Missions 1 and 4 (i.e., during acquisition) and decline between Missions 4 and 5 (i.e., when high workload sets in). The general finding that critical incident process behaviors change over missions suggests, as in Experiment 1, that these behaviors are highly malleable. Thus these behaviors can be changed rather quickly to suit a specific purpose. On the negative side, since these behaviors can change rather quickly across missions, regimented process behaviors may easily be lost under increased workload. The lack of a condition effect, unlike Experiment 1, is probably due in part to distributed critical incident process in Missions 1 and 2. As can be seen in Figure 19, unlike Experiment 1, distributed teams actually have comparable critical incident process at the outset relative to co-located teams. This inconsistency between experiments does not hold for the later missions in which co-located critical incident process is higher than distributed as in Experiment 1. Although this difference was not found to be significant across Missions 4 and 5, Figure 19 suggests that there was some difference in this manner.

Although we found no significant condition effects we performed a follow up similar to that performed for Experiment 1 critical incident process. We fit a discriminant analysis model using the component critical incident items to classify experimental conditions (co-located = 0, distributed =1). As in Experiment 1, we were interested in identifying which, if any, critical incident process items were especially good at classifying co-located vs. distributed teams that were consistent across experiments. These results are given in Table 41.

Table 41
Results of Discriminant Analysis

Process Item	Wilks' Lambda	*F*	df_{num}	df_{den}	Sig.	Standardized Weights
1	.991	.82	1	89	.37	-.10
2	.93	7.05	1	89	.01	-.28
3	.96	3.84	1	89	.05	.42
4	1.00	.02	1	89	.88	-.04
5	.96	3.43	1	89	.07	-.21
6	.73	32.28	1	89	.00	.86

Again, Item 6 was the big discriminator. To recap, this item asks whether the team discusses and assesses their team performance after their mission. As in Experiment 1, Experiment 2 co-located teams tended to do this while distributed teams did not. Another item that was consistently a good discriminator was Item 3, which asks whether or not teams explicitly discuss emergent mission parameters. Again, co-located teams tended to do this while distributed teams did not. No other items were found to be consistent across experiments. Therefore across the two experiments, critical incident items involving planning (Item 6) and adaptive process behaviors (Item 3) were found to be consistently good classifiers of whether a team was co-located or distributed. Again, although these particular process behaviors probably do not map on to performance differences in the UAV synthetic task, these results suggest some fundamental behavioral differences between co-located and distributed teams.

Summary process. As we had done in Experiment 1, we sought to compute an overall summary process score based on the four dimensions. Before combining these into an average we analyzed the summary dimensions using hierarchical centroid clustering using the 1-correlation metric in order to identify any high level dimensions among the individual summary dimensions that should be considered in computing an overall score. As depicted in Figure 20, differences in the distances between successive centroids were approximately linear, suggesting no strong clustering among the individual summary process dimensions. We therefore computed the arithmetic average of the four un-weighted summary process dimensions in order to obtain an overall summary process score for each mission. Descriptive statistics for summary process for co-located and distributed teams at each mission are presented in Table 42

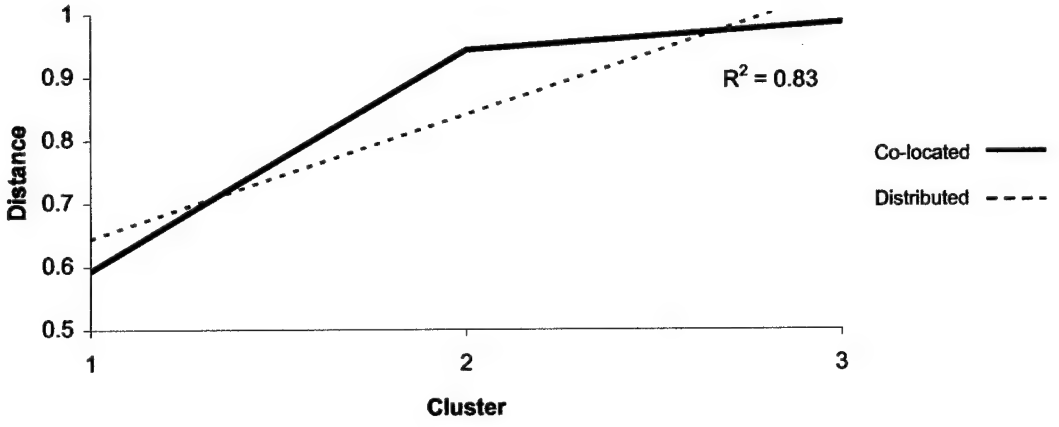


Figure 20. Experiment 2 summary process items; distance by cluster.

Table 42
Team Summary Process Scores for Co-located and Distributed Conditions

Mission	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
1	2.74	2.89	.84	1.07	1.75	1.38	4.13	4.50
2	2.97	3.36	.84	.92	2.00	2.13	4.13	4.88
3	3.51	3.80	.81	.73	2.25	3.00	5.00	5.00
4	3.70	3.92	1.11	.86	1.75	2.13	5.00	5.00
5	3.29	3.34	1.13	.91	1.25	1.63	4.38	4.75

As in Experiment 1, the planned analyses for summary process were identical to those for critical incident process. Therefore we tested for main effects of mission and condition and the interaction between the two, followed by interaction contrasts for acquisition (Missions 1 and 4) and workload (Missions 4 and 5). There were 100 total observations.

In the omnibus model, the main effect of mission was significant, $F(4, 72) = 9.31, p < .01$, indicating that over missions summary process scores changed (see Figure 21). The main effect of condition however, was not significant, $F(1, 18) < 1$. Thus it is not likely that the differences between co-located and distributed summary process depicted in Figure 21 are highly

dependable. Also, the omnibus interaction effect was not significant, $F(4, 72) < 1$, therefore mission differences were similar for both co-located and distributed teams.

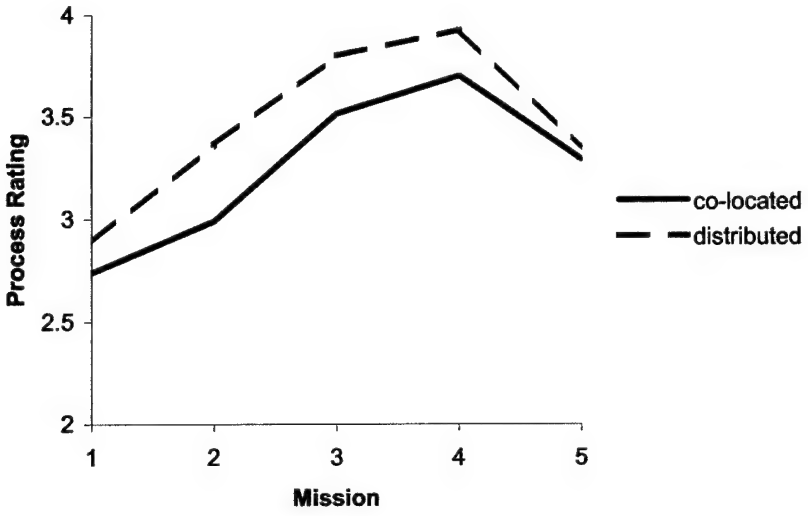


Figure 21. Mean Experiment 2 co-located and distributed summary process scores over missions.

The planned acquisition comparison main effect of mission revealed that teams in both conditions improved in summary process scores between Missions 1 and 4, $F(1, 18) = 14.51$, $p < .01$; see Figure 21. However, the condition main effect was not significant, $F(1, 18) < 1$. The averages of summary process across these two levels of mission therefore are similar across experimental conditions. The acquisition interaction effect was also not significant, $F(1, 18) < 1$, indicating that the observed change in summary process between the two missions in the acquisition comparison was similar for both co-located and distributed teams.

The workload planned comparison also revealed a significant mission effect, $F(1, 18) = 19.83$, $p < .01$, with teams in both conditions experiencing a sharp decline once high workload sets in at Mission 5 (see Figure 21). As in all other analyses, the main effect of condition for the workload comparison was not significant, $F(1, 18) < 1$. Apparently across the two "workload" missions, experimenters' ratings of team process were consistent for both co-located and distributed teams. And finally, the workload condition by mission interaction was also not significant, $F(1, 18) < 1$, indicating that the Mission 5 decrease was similar in nature across the distributed and co-located conditions.

The results from the planned analyses of summary process were largely consistent with those from Experiment 1. In general, co-located and distributed teams do not differ in terms of experimenters' ratings of overall quality of process behaviors. However these ratings tend to change over missions. Specifically, over Missions 1 to 4 we again find evidence of acquisition in terms of increasing quality of process behaviors. As we did for Experiment 1, we theorize that this supports the notion that over the first four missions, part of what teams are acquiring is the ability to coordinate with each other using high quality process behaviors and that this maps directly onto the performance acquisition curve. Also consistent with Experiment 1 we found

that the quality of the process behaviors in terms of summary process tend to decline when high workload sets in at Mission 5. We believe that increased workload has the tendency to impose constraints on the dynamics of the team that are not apparent in a low workload environment.

To summarize the most interesting findings:

- Although no differences in critical incident process between co-located and distributed teams were statistically detectable as in Experiment 1, we found support for the previous finding that planning and adaptive process behaviors typify co-located teams while distributed teams are less likely to exhibit these types of behaviors; such behaviors are not *absolutely* necessary to complete the mission

- Also consistent with Experiment 1, we found support for the notion that at least part of what all teams acquire while approaching performance asymptote is the ability to coordinate using high quality process behaviors

4.7.3 Situation Awareness

The analyses of situation awareness accuracy, similarity, and holistic accuracy are similar to the analyses on situation awareness in Experiment 1. Again the effects of interest include condition (co-located/distributed), mission, and type of query (repeated/non-repeated). The rationale for observing situation awareness for each type of query, rather than as an aggregate of repeated and non-repeated queries, is the same here as described in Experiment 1. That is, the repeated query seems to measure awareness of the experimental situation while the non-repeated queries seem to measure awareness of the task situation.

Situation awareness accuracy. Table 43 shows *situation awareness accuracy* on the repeated query and non-repeated queries for co-located and distributed teams on a mission-by-mission basis as well as averaged across the low workload missions. The accuracy score for a single team was missing for the non-repeated query at Mission 4. Therefore, the mean of the other 19 teams' accuracy for Mission 4 was used to replace the missing data point prior to calculating the overall mean for that mission.

A univariate, repeated measures analysis of variance with two repeated factors (mission and query type) and one between-subjects factor (condition) was used to analyze situation awareness accuracy. Results from the omnibus test are presented first, followed by the results from a series of contrasts aimed at answering more specific questions. A significant effect of mission was found, $F(4, 72) = 6.67, p < .01$, indicating that accuracy changed significantly over the course of the five missions. A main effect of query type was also found, $F(1, 18) = 24.98, p < .01$, where accuracy on non-repeated queries was significantly higher than accuracy on the repeated query. There was no main effect of condition, $F(1,18) = 1.66$. As can be seen in Figures 22 and 23, accuracy tended to change across missions differently for the repeated and non-repeated queries. This interaction was significant, $F(4, 72) = 11.43, p < .01$. The remaining two-way interactions were not significant. Specifically, query type did not interact with condition, $F(1, 18) < 1$, and mission did not interact with condition, $F(4, 72) < 1$. Finally, the three-way interaction between mission, query type, and condition also was not significant, $F(4, 72) = 1.51$.

Table 43

Situation Awareness Accuracy on the Repeated Query and Non-Repeated Queries for Co-located and Distributed Teams

Mission	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Repeated Query								
1 (LW)	.40	.80	.70	.79	.00	.00	2.00	2.00
2 (LW)	.60	.50	1.08	.71	.00	.00	3.00	2.00
3 (LW)	1.10	1.70	1.29	1.25	.00	.00	3.00	3.00
4 (LW)	2.20	2.30	1.32	.82	.00	1.00	3.00	3.00
5 (HW)	.60	.20	.84	.42	.00	.00	2.00	1.00
Average of Low Workload Missions	1.08	1.33	.77	.65	.00	.25	2.25	2.50
Non-Repeated Query								
1 (LW)	1.60	1.70	.84	1.06	.00	.00	3.00	3.00
2 (LW)	2.20	2.10	.92	1.20	.00	.00	3.00	3.00
3 (LW)	1.10	1.20	1.10	1.03	.00	.00	3.00	3.00
4 (LW)	1.81*	2.30	.79	1.06	.00	.00	3.00	3.00
5 (HW)	1.60	2.40	.84	.97	.00	.00	3.00	3.00
Average of Low Workload Missions	1.68	1.83	.50	.39	.75	1.25	2.51	2.25

* Contained missing data for one team, which was replaced with the mission mean

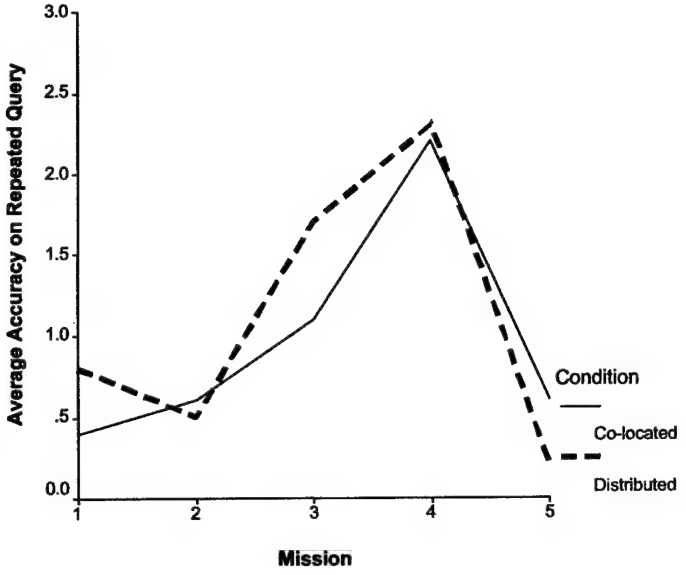


Figure 22. Situation awareness accuracy on the repeated query for co-located and distributed teams at each mission.

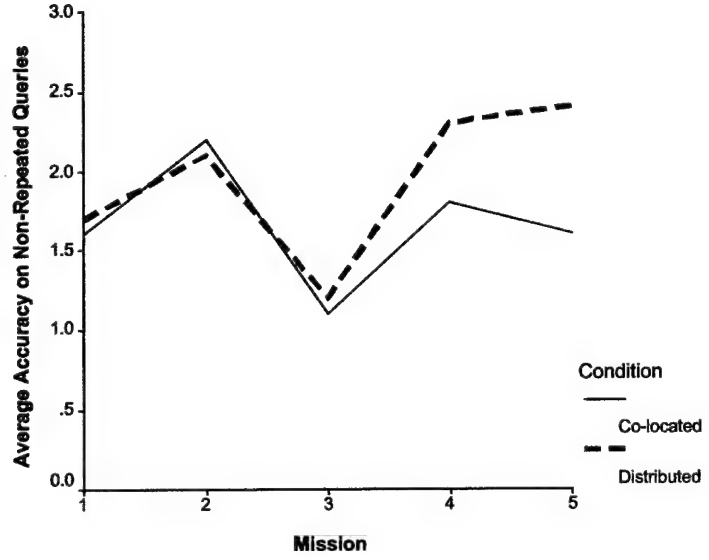


Figure 23. Situation awareness accuracy on the non-repeated queries for co-located and distributed teams at each mission.

Post hoc comparisons were conducted to locate the source of the mission by query type interaction. The comparisons revealed that accuracy on the non-repeated queries were significantly higher than accuracy on the repeated queries at Mission 1, Mission 2, and Mission 5 (see Table 2 for *t* statistics and *p*-values).

Table 44
T Statistics for the Comparison of the Average Accuracy on the Repeated Query Minus Average Accuracy on the Non-Repeated Queries at each Mission

Mission	*t* statistic	*p*-value
1	-4.47	.00
2	-5.29	.00
3	.67	.51
4	.71	.48
5	-6.84	.00

df = 19

A series of planned contrasts were also conducted in order to answer the following questions: (1) Did teams' accuracy improve over the low workload missions (Mission 1 vs. Mission 4), and (2) was there an effect of workload on accuracy (Mission 4 vs. Mission 5)? Although we were initially interested in the effect of condition (co-located/distributed) in answering these questions, condition was not included in the following contrasts since it was not found to be significant in the omnibus test. Univariate repeated measures analysis of variance with two repeated factors (mission and query type) were used to analyze each of the contrasts.

First, did teams' accuracy improve over the low workload missions? Accuracy did improve from Mission 1 to Mission 4, $F(1, 19) = 22.53, p < .01$. In addition, a main effect of query type was found, $F(1, 19) = 6.81, p = .02$, where accuracy on the non-repeated queries was significantly higher. Finally, a significant interaction between mission and query type emerged, $F(1, 19) = 9.93, p < .01$. *Post hoc* comparisons of Mission 1 and Mission 4 for the repeated query and non-repeated queries indicated that accuracy on the repeated query significantly improved, $F(1, 19) = 33.86, p < .01$, from Mission 1 to Mission 4 while accuracy on the non-repeated queries did not significantly improve, $F(1, 19) = 1.77$.

Second, was there an effect of workload on accuracy? Accuracy at Mission 4 was compared to Mission 5 in order to answer this question. An effect of workload was found, $F(1, 19) = 15.33, p < .01$, where accuracy significantly declined when the high workload mission was introduced. Furthermore, there was a significant effect of query type, $F(1, 19) = 10.59, p < .01$, with teams reaching higher levels of accuracy on the non-repeated queries than the repeated query. A significant interaction between workload and query type was also present, $F(1, 19) = 41.93, p < .01$. *Post hoc* comparisons were used to determine the source of the interaction. A comparison of Mission 4 to Mission 5 for the repeated query indicated that accuracy significantly declined in high workload, $F(1, 19) = 45.55, p < .01$, whereas for non-repeated queries, accuracy did not differ across the levels of workload, $F(1, 19) < 1$.

Situation awareness intrateam similarity. Table 45 shows *situation awareness intrateam similarity* on the repeated query for co-located and distributed teams on a mission-by-mission basis as well as for the average over low workload missions.

As with Experiment 1, data for intrateam similarity on the non-repeated queries were occasionally missing due to the fact that the truth of the query could change while administering the query. As a result, team members' responses were necessarily different in order to be accurate. Thus, intrateam similarity was not calculated in these cases. Instead, missing data at a particular mission were replaced with the mean for that mission. Of the 100 total missions (20 teams each with 5 missions), missing data were replaced for 7 missions.

Table 45
Situation Awareness Intrateam Similarity on the Repeated Query and Non-Repeated Queries for Co-located and Distributed Teams

Mission	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Repeated Query								
1 (LW)	.60	.50	.52	.53	.00	.00	1.00	1.00
2 (LW)	1.70	.50	1.16	.53	.00	.00	3.00	1.00
3 (LW)	1.90	1.70	1.20	1.42	.00	.00	3.00	3.00
4 (LW)	2.30	1.90	1.16	1.20	.00	.00	3.00	3.00
5 (HW)	.30	.30	.48	.48	.00	.00	1.00	1.00
Average of Low Workload Missions	1.63	1.15	.59	.67	.75	.00	2.50	2.00
Non-Repeated Query								
1 (LW)	1.34*	1.50	.94	1.08	.00	.00	3.00	3.00
2 (LW)	1.89*	1.89*	.99	1.29	1.00	.00	3.00	3.00
3 (LW)	1.32***	.80	.94	.92	.00	.00	3.00	3.00
4 (LW)	1.48*	2.20	.84	1.03	1.00	1.00	3.00	3.00
5 (HW)	1.00	2.10	.82	1.20	.00	.00	3.00	3.00
Average of Low Workload Missions	1.51	1.60	.52	.38	1.00	1.00	2.52	2.25

* Contained missing data for one team, which was replaced with the mission mean

A univariate, repeated measures analysis of variance with two repeated factors (mission and query type) and one between-subjects factor (condition) was used to analyze situation awareness intrateam similarity. The results are organized as they were for the analyses on accuracy, with the results from the omnibus test presented first, followed by the results from two planned contrasts. An effect of condition was not present, $F(1, 18) < 1$. However, a main effect of mission emerged, $F(4, 72) = 8.41, p < .01$, which confirmed that intrateam similarity changed significantly across missions (see Figures 24 and 25). There was also a significant main effect of query type on intrateam similarity, $F(1, 18) = 5.72, p = .03$, where teams' responses to the non-repeated queries were more similar than their responses to the repeated query.

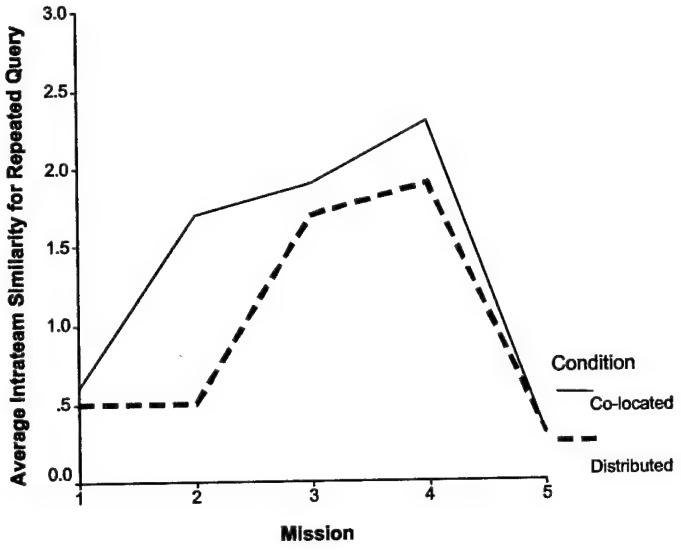


Figure 24. Average situation awareness intrateam similarity on the repeated query for both co-located and distributed teams at each mission.

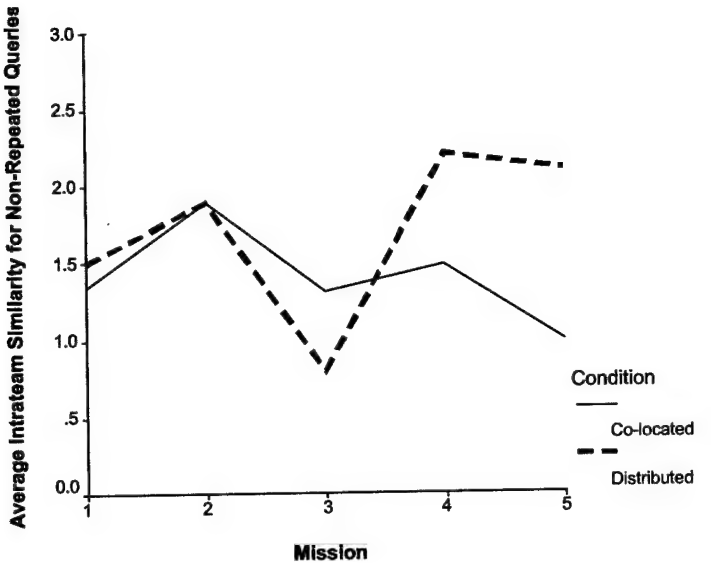


Figure 25. Average situation awareness intrateam similarity on the non-repeated queries for both co-located and distributed teams at each mission.

As with accuracy, intrateam similarity seemed to fluctuate across missions differently for the repeated query and the non-repeated queries. A test for this interaction between mission and query type was significant, $F(4, 72) = 7.40$, $p < .01$. *Post hoc* comparisons revealed that intrateam similarity differed significantly on the repeated and non-repeated queries for Mission 1, $t(19) = -3.55$, $p < .01$, Mission 2, $t(19) = -2.03$, $p = .06$, Mission 3, $t(19) = 2.00$, $p = .06$, and Mission 5, $t(19) = -4.80$, $p < .01$. Interestingly, teams were more similar in responding to the non-repeated queries during Mission 1, Mission 2, and Mission 5. However, during Mission 3, intrateam similarity scores on the repeated query significantly exceeded the scores on the non-repeated queries.

As Figure 26 illustrates, the interaction between condition and query type was also significant, $F(1, 18) = 4.41$, $p = .05$. A *post hoc* paired sample t-test indicated that the difference between (a) the mean difference between the co-located and distributed teams for the repeated query (1.4 – 1.0), and (b) the mean difference between co-located and distributed teams for the non-repeated queries (1.4 – 1.7) was significantly different from zero, $t(49) = 2.46$, $p = .02$. That is, although co-located and distributed teams did not differ much within a query type, the pattern of differences between co-located and distributed teams is reversed from the repeated query to the non-repeated query.

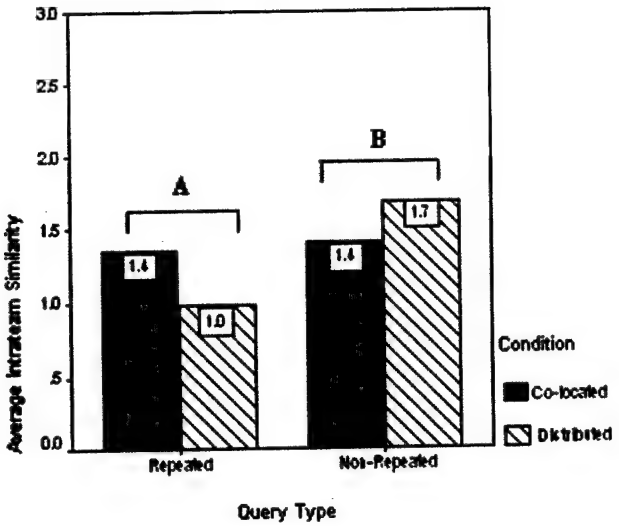


Figure 26. Average situation awareness intrateam similarity for the co-located and distributed teams on the repeated and non-repeated queries.

The interaction between mission and condition was also significant, $F(4, 72) = 2.34$, $p = .06$, which suggests that the effect of mission on intrateam similarity was moderated by condition. Collapsing across type of query, co-located teams appeared to be more similar than distributed teams through Mission 4 and then became less similar than distributed teams during Mission 5 (see Figure 27). *Post hoc* comparisons showed that for Mission 2, co-located teams were significantly more similar than distributed teams (see Table 46 for *t*-values). However, during Mission 5, distributed teams were significantly more similar than co-located teams. The three-way interaction among mission, condition, and query type was not significant, $F(4, 72) = 1.18$.

Table 46
Differences in Means of Situation Awareness Intrateam Similarity between Co-located Minus Distributed Teams at each Mission

Mission	Mean Difference between Co-located and Distributed	*p*-value
1	-.03	.91
2	.60	.04
3	.36	.31
4	-.16	.65
5	-.55	.06

df = 18

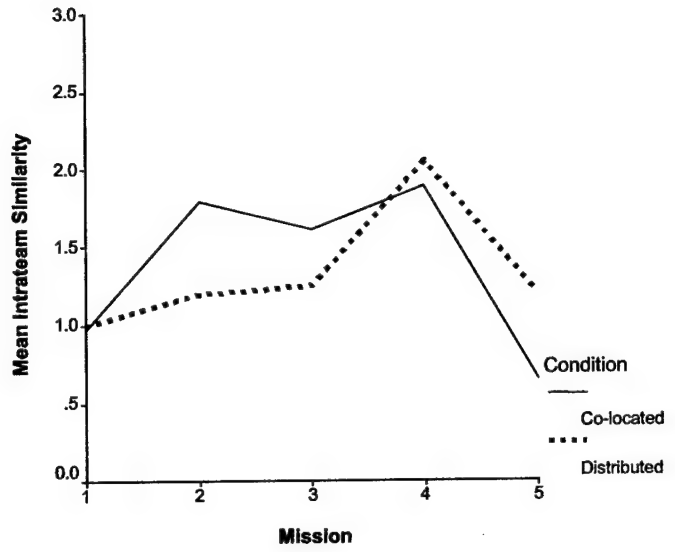


Figure 27. Average situation awareness intrateam similarity for co-located and distributed teams at each mission.

Two planned contrasts were conducted to further analyze intrateam similarity. The same questions that were answered for accuracy were answered here. Namely: (1) Did teams' intrateam similarity improve over the low workload missions (1 vs. 4), and (2) Was there an effect of workload on intrateam similarity (4 vs. 5)? Univariate, repeated measures analysis of variance with two repeated factors (mission and query type) and one between-subjects factor (condition) were used to analyze each of the contrasts. The effect of condition was included in the following contrasts since each two-way interaction involving condition was significant as reported in the omnibus test.

First, did co-located and distributed teams become more similar on their responses to the situation awareness queries from Mission 1 to Mission 4? There was no main effect of condition on similarity, $F(1, 18) < 1$, nor was there an effect of query type, $F(1, 18) = 2.16$. However, a main effect of mission was found, $F(1, 18) = 22.31, p < .01$, with teams reaching higher levels of intrateam similarity during Mission 4 than Mission 1. There was no significant interaction between condition and mission, $F(1, 18) < 1$, or between condition and query type, $F(1, 18) = 2.71$, but a significant interaction did emerge between mission and query type, $F(1, 18) = 7.42, p = .01$. This implies that the change in intrateam similarity from Mission 1 to Mission 4 differed, depending on query type. The three-way interaction among mission, condition, and query type was not significant, $F(1, 18) = 1.07$. *Post hoc* comparisons revealed that the mission by query type interaction stemmed from the fact that teams became more similar on their responses to the repeated query from Mission 1 to Mission 4, $F(1, 19) = 31.54, p < .01$, whereas similarity did not significantly differ from Mission 1 to Mission 4 on the non-repeated queries, $F(1, 19) = 1.91$.

The second contrast answered whether there was an effect of workload on intrateam similarity. A main effect of condition was not present, $F(1, 18) = 2.78$, but there was an effect of workload, $F(1, 18) = 21.74, p < .01$. Teams were significantly more similar in their responses to the

situation awareness queries at Mission 4 than at Mission 5. Furthermore, teams achieved higher similarity scores for the non-repeated queries than for the repeated query, as indicated by a significant main effect of query type, $F(1, 18) = 4.42$, $p = .05$. A significant interaction was also found between workload and query type, $F(1, 18) = 21.07$, $p < .01$, which demonstrates that intrateam similarity differed for the repeated and non-repeated queries over changing workload. There was also a significant interaction between condition and query type, $F(1, 18) = 5.52$, $p = .03$, which suggests that the effect of co-located/distributed status on intrateam similarity also depended on whether the queries were repeated or non-repeated. The condition by workload interaction was not significant, $F(1, 18) < 1$, nor was the interaction among condition, workload, and query type, $F(1, 18) < 1$.

Post hoc comparisons of Mission 4 and Mission 5 were performed separately for the repeated query and non-repeated queries in order to pin-point the source of the significant interactions. For the repeated query, a main effect of condition did not emerge, $F(1, 18) < 1$. However across conditions, intrateam similarity on the repeated query did significantly decrease during the high workload mission, $F(1, 18) = 44.18$, $p < .01$. There was no workload by condition interaction, $F(1, 18) < 1$. For the non-repeated query, intrateam similarity was not affected by workload, $F(1, 18) = 1.05$, nor did workload interact with condition, $F(1, 18) < 1$. However, a significant main effect of condition was found, $F(1, 18) = 7.34$, $p = .01$, where distributed teams were more similar in their responses on the non-repeated queries than co-located teams.

Holistic situation awareness accuracy. Table 47 shows *holistic situation awareness accuracy* descriptive statistics for co-located and distributed teams on a mission-by-mission basis. The table also shows an average of holistic accuracy over the low workload missions. As with accuracy, the holistic accuracy score for a single team was missing for the non-repeated query at Mission 4. Therefore, the mean of the other 19 teams' holistic accuracy at Mission 4 was used to replace the missing data point prior to calculating the overall mean for that mission.

A univariate, repeated measures analysis of variance with two repeated factors (mission and query type) and one between-subjects factor (condition) was used to analyze situation awareness holistic accuracy. Results from the omnibus test are presented first, followed by the results from two planned contrasts. First, there was no main effect of condition on holistic accuracy, $F(1, 18) < 1$. As with accuracy and intrateam similarity, a main effect of mission was found for holistic accuracy, $F(4, 72) = 6.07$, $p < .01$. A significant main effect of query type was also found, $F(1, 18) = 89.80$, $p < .01$, where holistic accuracy was significantly higher for non-repeated queries than for the repeated query. Figures 28 and 29 illustrate the different trends in holistic accuracy on the repeated and non-repeated queries, respectively. The interaction between mission and query type was significant, $F(4, 72) = 6.85$, $p < .01$, which indicates that the effect of mission on holistic accuracy depended on whether the queries were repeated or non-repeated. The interaction between condition and query type was not significant, $F(1, 18) = 1.25$, nor was the interaction between mission and condition, $F(4, 72) < 1$. Finally, the three-way interaction among mission, condition, and query type was also not significant, $F(4, 72) < 1$.

Table 47

Situation Awareness Holistic Accuracy on the Repeated Query and Non-Repeated Queries for Co-located and Distributed Teams

Mission	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Repeated Query								
1 (LW)	.10	.10	.32	.32	.00	.00	1.00	1.00
2 (LW)	.10	.30	.32	.48	.00	.00	1.00	1.00
3 (LW)	.40	.70	.52	.48	.00	.00	1.00	1.00
4 (LW)	.70	.80	.48	.42	.00	.00	1.00	1.00
5 (HW)	.10	.10	.32	.32	.00	.00	1.00	1.00
Average of Low Workload Missions	.33	.48	.29	.30	.00	.00	1.00	1.00
Non-Repeated Query								
1 (LW)	.80	.90	.42	.32	.00	.00	1.00	1.00
2 (LW)	1.00	.70	.00	.48	1.00	.00	1.00	1.00
3 (LW)	.60	.80	.52	.42	.00	.00	1.00	1.00
4 (LW)	.89*	.90	.31	.32	.00	.00	1.00	1.00
5 (HW)	.90	.90	.32	.32	.00	.00	1.00	1.00
Average of Low Workload Missions	.82	.83	.17	.12	.50	.75	1.00	1.00

* Contained missing data for one team, which was replaced with the mission mean

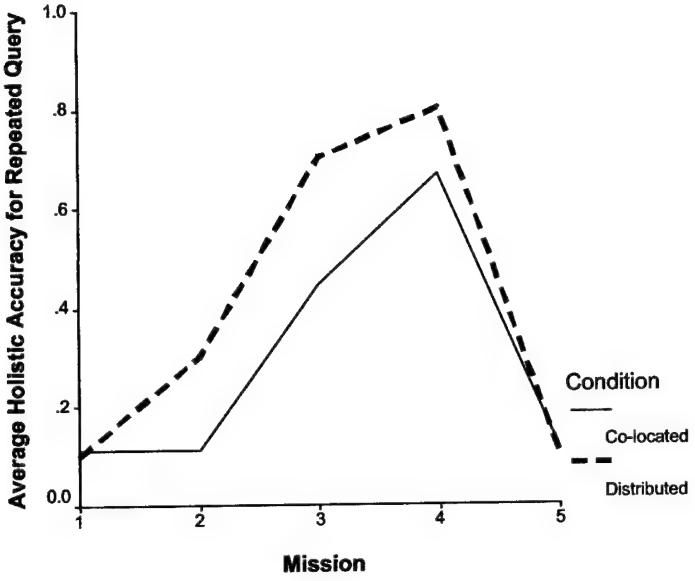


Figure 28. Average situation awareness holistic accuracy on the repeated query for both co-located and distributed teams at each mission.

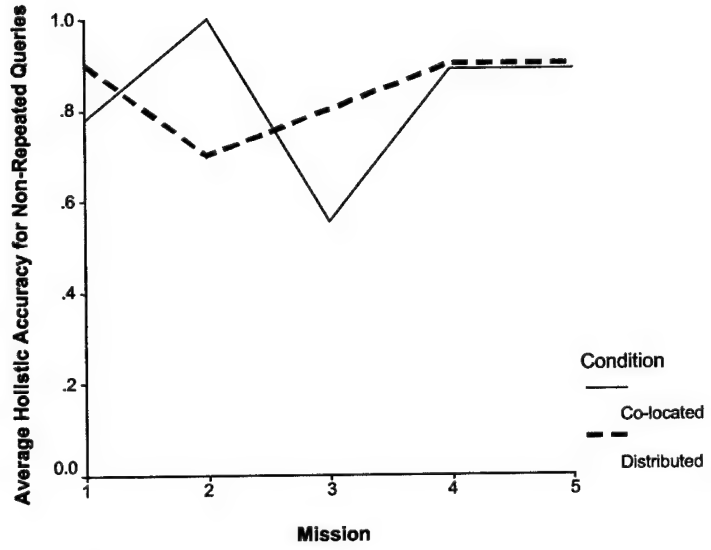


Figure 29. Average situation awareness holistic accuracy on the non-repeated queries for both co-located and distributed teams at each mission.

Post hoc comparisons were conducted in order to explain the mission by query type interaction. As with accuracy, the comparisons revealed that holistic accuracy on the non-repeated queries was significantly higher than holistic accuracy on the repeated queries at Mission 1, Mission 2, and Mission 5 (see Table 48 for *t* statistics and *p*-values).

Table 48
T Statistics for the Comparison of the Average Holistic Accuracy on the Repeated Query Minus Average Holistic Accuracy on the Non-Repeated Queries at each Mission

Mission	*t* statistic	*p*-value
1	-7.55	.00
2	-4.95	.00
3	-1.00	.33
4	-1.31	.20
5	-6.84	.00

df = 19

Planned contrasts were conducted in order to answer (1) whether teams' holistic accuracy improved over the low workload missions (1 vs. 4), and (2) whether there was an effect of workload on holistic accuracy (4 vs. 5). As with accuracy, we were initially interested in the effects of the co-located/distributed status, but the lack of significant effects in the omnibus test warranted the exclusion of the condition factor from further tests. Univariate repeated measures analysis of variance with two repeated factors (mission and query type) were used to analyze each of the contrasts.

First, did holistic accuracy improve from Mission 1 to Mission 4? For both queries, teams' holistic responses were significantly more accurate in Mission 4 than in Mission 1, $F(1, 19) = 18.07, p < .01$. There was also a main effect of query type, $F(1, 19) = 39.03, p < .01$, where teams' holistic responses were more accurate on non-repeated queries than the repeated query. A significant interaction between mission and query type also surfaced, $F(1, 19) = 15.66, p < .01$. *Post hoc* comparisons revealed that for the repeated queries, holistic accuracy improved significantly when comparing Mission 1 to Mission 4, $F(1, 19) = 35.29, p < .01$, but holistic accuracy did not improve from Mission 1 to Mission 4 for non-repeated queries, $F(1, 19) < 1$.

The second contrast compared Mission 4 to Mission 5 in order to address whether there was an effect of workload on holistic accuracy. Holistic accuracy scores were significantly lower in Mission 5 than in Mission 4, $F(1, 19) = 15.06, p < .01$. Accuracy of holistic responses was also effected by query type, $F(1, 19) = 68.48, p < .01$, where teams were more accurate in their holistic responses to the non-repeated queries than to the repeated query. Moreover, mission and query type interacted significantly, $F(1, 19) = 11.15$. Post hoc comparisons were used to examine the source of this interaction. Specifically, Mission 4 was compared to Mission 5 for the repeated query and non-repeated queries separately. Workload did effect teams' holistic accuracy on the repeated query, $F(1, 19) = 18.78, p = .01$, where holistic accuracy was higher during low workload (i.e., Mission 4) than in high workload (i.e., Mission 5). In contrast, holistic accuracy on the non-repeated queries did not change significantly from low workload to high workload, $F(1, 19) < 1$.

To summarize:

- For situation awareness accuracy measures (accuracy, holistic accuracy), there was no effect of dispersion condition.
- Situation awareness similarity within teams was greater for co-located teams for the first four missions and was greater for distributed teams at Mission 5. Also, co-located teams tended to be more similar on the repeated queries whereas distributed teams were more similar on the nonrepeated queries.
- Accuracy, similarity, and holistic accuracy improved between Missions 1 and 4 for repeated queries, but not for nonrepeated queries.
- Accuracy and holistic accuracy declined between Missions 4 and 5 for repeated queries, but not for nonrepeated queries

4.7.4 Taskwork Knowledge

The means and standard deviations as well as the minimum and maximum scores for *overall taskwork accuracy* can be seen in Table 49 for distributed and co-located teams. The means reveal that co-located teams were more accurate. A one-way analysis of variance (ANOVA) revealed that this difference was significant, $F(1, 18) = 6.42, p = .02$.

Table 49
Taskwork Accuracy in Co-located and Distributed Conditions

Mean		Standard Deviation		Minimum		Maximum	
Col	Dist	Col	Dist	Col	Dist	Col	Dist
.50	.44	.04	.06	.45	.35	.59	.52

Table 50 displays the descriptive statistics for *taskwork positional knowledge*. The means reveal that team members in the co-located condition had more knowledge about their own roles than did members of distributed teams. A one-way ANOVA confirmed this finding with a significant main effect of condition, $F(1, 18) = 6.21, p = .02$.

Table 50
Taskwork Positional Knowledge in Co-located and Distributed Conditions

Mean		Standard Deviation		Minimum		Maximum	
Col	Dist	Col	Dist	Col	Dist	Col	Dist
.30	-.30	.57	.52	-.45	-1.1	1.3	.72

Taskwork interpositional knowledge was also analyzed as a function of the co-located/distributed manipulation. As with positional (role) knowledge, there was a significant main effect for condition, $F(1, 18) = 4.31, p = .05$, indicating that co-located teams had more knowledge about other team members' roles.

Table 51
Taskwork Inter-Positional Knowledge in Co-located and Distributed Conditions

Mean		Standard Deviation		Minimum		Maximum	
Col	Dist	Col	Dist	Col	Dist	Col	Dist
.21	-.21	.84	.42	-.52	-.84	.84	.49

Taskwork intrateam similarity descriptive data are shown in Table 52. There was a significant main effect for condition, $F(1, 18) = 5.18$, $p < .05$, with co-located teams being more similar in terms of taskwork knowledge than distributed teams.

Table 52
Taskwork Similarity in Co-located and Distributed Conditions

Mean		Standard Deviation		Minimum		Maximum	
Col	Dist	Col	Dist	Col	Dist	Col	Dist
.41	.33	.09	.07	.30	.19	.57	.47

The final taskwork variable examined was *holistic taskwork accuracy*. Descriptive data are displayed in Table 53. This variable revealed no significant main effect for condition, $F(1, 18) < 1$.

Table 53
Taskwork Holistic Accuracy in Co-located and Distributed Conditions

Mean		Standard Deviation		Minimum		Maximum	
Col	Dist	Col	Dist	Col	Dist	Col	Dist
.61	.62	.13	.13	.48	.48	.86	.86

To summarize: Co-located teams obtained higher scores on all of the taskwork variables with the exception of holistic taskwork accuracy for which there was no significant difference. Interestingly, holistic taskwork accuracy involves a consensus-building component, which is not present in the other taskwork measures.

4.7.5 Teamwork Knowledge

The means and standard deviations as well as the minimum and maximum scores for overall teamwork accuracy can be seen in Table 54 for co-located and distributed teams. The means reveal that co-located teams scored slightly better than distributed teams. However, a univariate ANOVA revealed no significant effect of condition, $F(1, 18) < 1$.

Table 54
Overall Teamwork Accuracy in Co-located and Distributed Conditions

Mean		Standard Deviation		Minimum		Maximum	
Col	Dist	Col	Dist	Col	Dist	Col	Dist
23.47	22.90	1.80	3.51	21.33	17.00	26.67	28.67

Knowledge of one's own role (AVO, PLO, or DEMPC), *positional knowledge*, as well as knowledge of other roles, *inter-positional knowledge*, were also examined for teamwork. Descriptive statistics for these variables are displayed in Tables 55 and 56.

Table 55
Teamwork Positional Knowledge in Co-located and Distributed Conditions

Mean		Standard Deviation		Minimum		Maximum	
Col	Dist	Col	Dist	Col	Dist	Col	Dist
.56	.56	.06	.09	.47	.43	.67	.70

Table 56
Teamwork Inter-Positional Knowledge in Co-located and Distributed Conditions

Mean		Standard Deviation		Minimum		Maximum	
Col	Dist	Col	Dist	Col	Dist	Col	Dist
.54	.50	.07	.11	.43	.35	.65	.70

Values are based on percentage correct because the number of items on which a score was based varied by role. There are no significant differences between co-located and distributed teams on these variables. A univariate ANOVA revealed there was no main effect of condition for positional, $F(1, 18) < 1$, or interpositional, $F(1, 18) < 1$, knowledge. As can be seen in Table 57, *intrateam similarity* was higher for distributed teams than co-located teams, $F(1, 18) = 3.92$, $p = .06$.

Table 57
Teamwork Similarity in Co-located and Distributed Conditions

Mean		Standard Deviation		Minimum		Maximum	
Col	Dist	Col	Dist	Col	Dist	Col	Dist
7.10	9.00	1.52	2.62	5.00	4.00	10.00	13.00

Holistic teamwork accuracy was also investigated to determine whether co-located and distributed teams differed on this variable. The means in Table 58 show that co-located teams scored slightly better than distributed teams. However, the effect of condition was not significant, $F(1, 18) < 1$.

Table 58
Holistic Teamwork Accuracy in Co-located and Distributed Conditions

Mean		Standard Deviation		Minimum		Maximum	
Col	Dist	Col	Dist	Col	Dist	Col	Dist
26.80	25.20	4.16	4.13	17.00	19.00	31.00	32.00

To summarize:

- Distributed teams were more similar in terms of teamwork knowledge than co-located teams
- However, no teamwork accuracy metric (overall accuracy, positional accuracy, interpositional accuracy, holistic accuracy) differed for co-located and distributed teams

4.7.6 Correlations of Performance and Process

Table 59 displays the correlations between team performance and process scores for Missions 4 and 5. Co-located and distributed teams were not examined separately because there was no effect of dispersion condition on either performance or process. Missions 4 and 5 were used to represent the low workload and high workload missions respectively due to significant changes during low workload missions on the team performance and process measures.

In the high workload mission, teams that obtained higher critical incident process scores also performed better as a team. Although this correlation over low workload mission (Mission 4) is in the same direction, it was not found to be significant. For both low workload and high workload missions, teams that received higher summary process scores tended to perform better on the UAV synthetic task.

Table 59
Correlations Between Performance and Process

	Critical Incident Process	Summary Process
Mission 4 (n = 20)	0.14	0.68**
Mission 5 (n = 20)	0.38*	0.59**
Overall (n = 100)	0.40**	0.66**

* $p < .05$. ** $p < .01$

To summarize:

- Overall, there was a significant positive correlation between both critical incident and summary process and team performance; teams with good process also tend to perform well

- Although both critical incident process and summary process are predictive of team performance, critical incident process was a better predictor of high workload performance, compared to low workload performance; summary process works well at both workload levels

4.7.7 Correlations Between Knowledge Measures and Performance or Process

In Experiment 2 there were 16 separate knowledge measures considered. Taskwork and teamwork (Knowledge Session 2 only) were scored against overall, positional, and interpositional referents, as well as for similarity, yielding a total of eight taskwork and teamwork measures. Situation awareness also involved a total of eight measures with four each for repeated and nonrepeated queries. The four included situation awareness accuracy and similarity each scored in low (Mission 4) and high (Mission 5) workload missions.

In order to summarize the correlations among the 16 knowledge measure variables, they were subjected to a hierarchical cluster analysis utilizing the centroid linkage method. Using Pearson correlations significant at $p \leq .10$ as a cluster cutoff, twelve variables formed six distinct, non-overlapping clusters. The remaining four factors did not enter into a cluster. Table 60 presents the clusters and the knowledge measures that form them.

Table 60
Clusters Among Knowledge Measures for Experiment 2

Cluster Name	Variables
1) Taskwork Accuracy-Positional	Taskwork Overall Accuracy Taskwork Positional Knowledge
2) Taskwork Similarity-IPK	Taskwork Similarity Taskwork Interpositional Knowledge
3) Teamwork	Teamwork Accuracy Teamwork Positional Knowledge
4) SA Non-Repeated Low Workload	SA Accuracy Non-Repeated Low Workload SA Similarity Non-Repeated Low Workload
5) SA Non-Repeated High Workload	SA Accuracy Non-Repeated High Workload SA Similarity Non-Repeated High Workload
6) SA Repeated Low Workload	SA Accuracy Repeated Low Workload SA Similarity Repeated Low Workload

Relationship among knowledge clusters and team performance. To correlate each cluster with team performance, the variables within each cluster were standardized (i.e., if not already to scale) and averaged. Correlations between the clusters and team performance as well as between the four single variables and team performance can be seen in Table 61. The large positive correlation between situation awareness for the repeated SA queries in low workload (Cluster 6) and both low workload and high workload performance indicates that good situation awareness (repeated query) was associated with higher performance. Of those knowledge variables that did not fit into a cluster, only situation awareness accuracy to the repeated query in high workload correlated significantly with high workload performance, in the positive direction.

Table 61
Correlations Between Knowledge Measures Clusters and Team Performance

Cluster/Variable	Low Workload Performance	High Workload Performance
Cluster 1 - Taskwork Accuracy-Positional	.27	.37
Cluster 2 -Taskwork Similarity-IPK	.13	.13
Cluster 3 - Teamwork	-.10	.16
Cluster 4 - SA Non-Repeated Low Workload	-.09	-.16
Cluster 5 - SA Non-Repeated High Workload	-.34	-.25
Cluster 6 - SA Repeated Low Workload	.60**	.58**
Teamwork Similarity	.31	.11
Teamwork IPK	.24	.08
SA Similarity Repeated High Workload	.04	.05
SA Accuracy Repeated High Workload	-.04	.42*

* $p < .05$. ** $p < .01$ $df = 18$

Although there are no linear relations between the taskwork and teamwork knowledge measures and team performance, there are some quadratic relations. Specifically, when controlling for the quadratic relationship between the teamwork cluster (Cluster 3) and high workload performance, the linear relationship between the two variables is significant, pr (17) = .49, p = .03. Also, when controlling for the quadratic relationship between teamwork similarity and low workload performance, the linear relationship is significant, pr (17) = .46, $p < .05$. And finally, when controlling for the quadratic relationship between teamwork similarity and high workload performance, the linear relationship is significant, pr (17) = .45, $p < .06$. These quadratic relations between teamwork knowledge and team performance suggest that teamwork knowledge at some middle level is associated with optimal performance. Too little teamwork knowledge may indicate poor understanding of the teamwork requirements of the task. On the other hand, too much teamwork knowledge may not only be unnecessary for optimal performance, but the acquisition of this higher level of knowledge may detract from the acquisition of other more important team skills like coordination.

Relationship between knowledge clusters and team process. As mentioned above, knowledge variables within each cluster were standardized and averaged in order to correlate each cluster with process. Correlations between the knowledge measures (clusters and single variables) and critical incident process are reported in Table 62. The SA non-repeated high workload cluster (Cluster 5) was found to be negatively associated with critical incident process during low workload and also during high workload for co-located teams. This suggests that teams with good process behaviors at critical times during the missions tended to exhibit poor situation awareness to the non-repeated queries in high workload. Another significant correlation emerged between teamwork similarity and distributed teams' critical incident process during high workload, which indicates that distributed teams made up of individuals who were more similar in their responses on the teamwork questionnaire also showed better process behaviors at critical mission incidents.

Table 62
Correlations Between Knowledge Measures Clusters and Critical Incident Process

Cluster/Variable	Low Workload CIP	High Workload CIP	
		Co-located	Distributed
Cluster 1 - Taskwork Accuracy-Positional	.12	-.18	.50
Cluster 2 -Taskwork Similarity-IPK	.05	-.22	.23
Cluster 3 - Teamwork	.05	-.33	.07
Cluster 4 - SA Non-Repeated Low Workload	.16	.19	.16
Cluster 5 - SA Non-Repeated High Workload	-.47*	-.63*	-.47
Cluster 6 - SA Repeated Low Workload	.25	.30	.32
Teamwork Similarity	-.12	.25	.69*
Teamwork IPK	.13	.41	.54
SA Similarity Repeated High Workload	-.09	-.37	-.50
SA Accuracy Repeated High Workload	-.14	-.38	.21

* $p \leq .05$. $df = 18$ low workload $df = 8$ high workload

Correlations between the knowledge measures (clusters and single variables) and summary process can be seen in Table 63. The situation awareness repeated low workload cluster (Cluster 6) was significantly correlated with summary process in both low workload and high workload missions, where teams with good situation awareness for the repeated query in low workload were also rated as demonstrating good process behaviors. Correlations between teamwork similarity and summary process at both low workload and high workload missions also indicate that teams with members who responded similarly on the teamwork questionnaire also received high ratings on their process behaviors.

Table 63
Correlations Between Knowledge Measures Clusters and Summary Process

Cluster/Variable	Low Workload Summary Process	High Workload Summary Process
Cluster 1 - Taskwork Accuracy-Positional	.22	.30
Cluster 2 -Taskwork Similarity-IPK	-.10	-.13
Cluster 3 - Teamwork	.31	.28
Cluster 4 - SA Non-Repeated Low Workload	.16	.12
Cluster 5 - SA Non-Repeated High Workload	-.23	-.36
Cluster 6 - SA Repeated Low Workload	.75**	.62**
Teamwork Similarity	.51*	.44*
Teamwork IPK	.22	.20
SA Similarity Repeated High Workload	.04	-.03
SA Accuracy Repeated High Workload	-.06	.02

*$p \leq .05$. ** $p < .01$. $df = 18$

To summarize:

- Situation awareness non-repeated high workload scores were negatively related to critical incident process, while situation awareness repeated low workload scores were positively related to summary process scores.

- For both types of process ratings, high teamwork similarity (i.e., distributed teams) was indicative of good team process.

4.8 Experiment 2: Discussion

This experiment was a replication of Experiment 1using all-male teams and some slight procedural changes. In this experiment the effect of co-located versus distributed mission environments on team performance, process, and cognition was investigated. The team task was a UAV reconnaissance task and involved three individuals who worked together in five 40-minute missions, the last one under higher workload than the first four. Each primary dependent measure was analyzed in order to address the four hypotheses raised previously. The results are summarized in Table 64 in terms of answers to three main questions: (1) Was dispersion detrimental, (2) Was there early improvement (i.e., learning), and 3) Was increased workload detrimental.

Table 64
Summary of Experiment 2 Results

Measure	Was Dispersion Detrimental?	Was There Early Improvement (i.e., Learning)?	Was Increased Workload Detrimental?
Team performance	No	Yes	Yes
Team Process	No	Yes	Yes
Situation Awareness	Yes for early missions, not for M5	Yes, repeated queries only	Yes, repeated queries only
Taskwork Knowledge	Yes	N/A	N/A
Teamwork Knowledge	Teamwork similarity is higher for distributed	N/A	N/A

Results regarding dispersion and performance found in Experiment 1 were replicated in Experiment 2. That is, geographic dispersion was not detrimental to team performance in our synthetic task environment. It was not detrimental to learning, nor in high workload environments. In fact, the slight distributed advantage seen in Experiment 1 under high workload was also seen, though not statistically significant, in the low workload missions of Experiment 2. Thus, the first hypothesis (H2.1) regarding poorer performance for distributed teams was not supported.

We also hypothesized that there would be process deficits associated with the distributed environment during task acquisition that would drive early performance deficits as well as knowledge and situation awareness deficits. This hypothesis was supported by the critical incident process measure in Experiment 1, but was not replicated in Experiment 2, though the means for critical incident process favor co-located teams in four out of five missions. However, the discriminant analysis of critical process items done in Experiment 2 replicated the results of Experiment 1. That is, the items that most distinguished co-located from distributed teams had to do with planning and adaptive behaviors, which the co-located teams seemed to carry out more readily than distributed teams.

Unlike Experiment 1 in which process, but not knowledge or situation awareness suffered from dispersion, in Experiment 2, process did not suffer from dispersion, rather knowledge and situation awareness were affected. Co-located teams had superior team situation awareness (in early missions) and had more taskwork knowledge (except for the holistic metric, in which distributed teams approached co-located accuracy). On the other hand, distributed teams had greater intrateam similarity when it came to teamwork knowledge. The fact that the manipulation affected knowledge without affecting process is interesting and may suggest that our process measures are limited in terms of sensitivity. Again we find only partial support for our hypothesis about process and knowledge deficits associated with dispersion (H2.2).

It is important to note that the changed placement of our knowledge session from the beginning and end of the experimental session to the middle of the session seems to have made a difference for the knowledge measures which have revealed some dispersion effects in Experiment 2. It may be the case that the changed placement of the elicitation session provided data that were more sensitive to these differences.

Further, increased workload did seem to have detrimental effects on performance, process, and situation awareness. For situation awareness (similarity) workload was generally detrimental, but more so for the co-located teams. In all other cases, workload affected both co-located and distributed teams similarly. Therefore it was not generally the case that distributed teams suffered more than co-located teams under high workload and thus our third hypothesis (H2.3) was also not supported.

Our fourth hypothesis (H2.4) concerned the contribution of individual differences among teams as a moderator of process and performance effects. In Experiment 2 we took some additional measures of individual attributes (verbal processing speed, voice stress) that should help us to answer this question. The results pertaining to these hypotheses are presented in the section on archival analyses of individual difference factors. In that section we examine the role of individual differences in team performance and cognition making use of data from Experiments 1 and 2 in addition to two previous experiments conducted in the CERTT Lab.
There are two other general findings that are worthy of note. The first concerns repeated versus nonrepeated forms of situation awareness queries. In almost all cases in Experiments 1 and 2, effects were found for repeated, but not non-repeated queries. Repeated queries seemed to mirror performance more directly which is also evident in the correlational data. However, we suspect that good teams may have developed good mission-specific strategies over time for estimating the number of targets that they would acquire (i.e., to respond to the repeated query).

This strategy was interfered with the introduction of the high workload mission in which total target number was changed. Thus, team situation awareness as measured by the repeated queries may tell us more about a team's ability to gauge their own performance on a repeated scenario, rather than the team's true situation awareness.

Another finding that is robust throughout both studies and measures is the finding that teams improve. In both experiments team performance, process, and situation awareness (repeated queries only) increased over the first four missions. In Experiment 1 in which there were two knowledge sessions, taskwork and teamwork knowledge also improved. Thus, because we see these kinds of learning effects in our data we conclude that our failure to find a condition effect is either because there is no effect or it is smaller than the learning and workload effects, which are readily detected in our studies.

In conclusion, we found in these first two studies that distributed mission environments as defined in the context of our UAV-STE have virtually no negative impact on team performance and may in fact, be beneficial in some regards. Specifically, in our later analysis of workload measures we describe a finding in which co-located DEMPCs perceive greater levels of workload than distributed DEMPCs suggesting that there may be some subtle social effects that can work in favor of dispersion. We also speculate that distributed teams may be forced to adapt to a different, possibly more structured style, of team interaction or team process and therefore, may be in a better position when workload demands necessitate efficient coordination.

These conclusions regarding distributed environments should be interpreted in light of the specific task required of the UAV-STE and in particular the fact that co-located teams communicated over head sets and therefore used an identical mode of communication to distributed teams. A true face-to-face environment may indeed have benefits over our distributed environment. However, our co-located environment mirrors the actual co-located environment of the Predator UAV in which operators sitting side by side communicate using microphones and headsets. Further, there were some negative effects of DMEs. Distributed teams tended to have poorer process when it came to planning and adaptive behaviors and lower levels of taskwork, but not teamwork knowledge. In the case of team process, we speculate that the distributed teams adapted to their situation by adopting a *different* set of interaction behaviors. Perhaps the omission of non-mission essential behaviors was necessary for more efficient interaction in this environment and may indeed be adaptive in high workload settings.

The fact that distributed teams have a poorer understanding of the task than co-located teams can be explained by the fact that they are impaired in their ability to observe the computer screens and behaviors of their fellow team members. This deficit may make it difficult to develop an understanding of the task from a "big picture" perspective or from the perspective of other team roles. However, this deficit did not seem to affect team performance. The fact that the teams in our task are trained to a criterion level of knowledge means that they may all have the taskwork knowledge that they need at the start of the missions. Understanding the task at a higher level is sometimes correlated with good team performance, but is not itself the most critical factor. Instead of taskwork knowledge, we suspect that one difference between good and poor teams is the ability to coordinate in an adaptive and timely fashion. This is what we consider to be cognitive processing at a team level. This is supported by correlations between process and team

performance. We believe that this team-level cognitive processing is what teams learn in the first few missions of an experiment and that this may distinguish good from poor teams at asymptote. We will explore the question of what is critical for effective team performance in the next three sections of this report. In the next section we describe a small study that we did under Objective 2 of this project to empirically benchmark team performance in our UAV-STE.

4.9. Experiment 3: Benchmarking Study

This experiment was carried out under Objective 2 of this project, which is to *conduct an empirical study to benchmark expert performance in the context of the CERTT Lab's three-person UAV-STE.* The three tasks associated with this objective are as follows:

- **Task 1:** Determine requirements for expert teams and recruit expert teams for experiment
- **Task 2:** In a single session with five missions collect performance, cognitive, and process data from expert teams
- **Task 3:** Compare data from expert teams to previously collected data from non-expert teams.

The purpose of this objective is to evaluate the validity and cognitive fidelity of the UAV synthetic task environment in a study that uses expert UAV operators. In addition, an empirically-derived performance benchmark will *set the standard* for future interventions in the lab designed to improve performance.

As mentioned in the discussion following Experiment 2, we also see this as an opportunity to identify factors most relevant to expertise in this UAV-STE. That is, how do teams who perform at expert levels compare to other teams in terms of process, situation awareness, and knowledge? Similarly, we see connections between the individual differences thrust of this project and this benchmarking experiment. Specifically, degree of experience working as a member of a UAV ground control team is an individual characteristic that is likely to differ among operators and play a significant role in predicting team performance and cognition.

The concept of benchmarking performance as a test of UAV-STE validity also requires some explanation. Much of our work has dealt with validity of the measures that we take in the lab context, however the validity of that context itself can also be questioned. That is, is the CERTT UAV-STE a valid test bed? Is it faithful to the field of practice? Thus far, we have addressed this issue in two ways. First, the fact that the task itself was based on a cognitive task analysis in the field of practice indicates that it is valid to the extent that the cognitive task analysis and our interpretation of that analysis are valid. Second, various UAV experts have judged the synthetic task in terms of its face validity. On both of these counts validity has been supported.

The proposed benchmarking study presents a third opportunity for validation. In this case we assume that to the extent that the STE is faithful to team cognition in the field of practice, then operators experienced in that field, should excel on the aspects of the synthetic task involving

team cognition. It is important to point out that we view validity or fidelity as multifaceted and that one simulation may be faithful to the look and feel of the task equipment, whereas another may be faithful to the cognitive processes required of the task. Thus, we do not expect expert UAV operators to excel on aspects of the synthetic task related to our custom interface, which is novel to them. We do, however, expect them to excel on aspects relevant to team cognition. The extent to which this occurs provides a test of validity and also sets a benchmark for improving novice performance through training or interface design. In a similar benchmarking study with a high fidelity simulation of a UAV ground operations station, Schreiber, Lyon, Martin, and Confer (2002) found that experienced Predator UAV operators performed consistently better than other groups including pilots trained on other aircraft. The authors interpreted this as evidence for the validity of their simulation in terms of the requisite stick-and-rudder skills. In the same way, we hope to validate the CERTT UAV-STE in terms of the team cognition required for effective task performance.

Based on previous experience and data collection in our UAV synthetic task environment, we anticipate that experienced operators will need to learn the new interface and may even have some negative transfer due to interface differences between our simulated and the actual ground stations. We therefore do not anticipate any advantage in early individual training on the task for experienced operators. Experience, however, should play a role in acquisition of teamwork skill and therefore performance, which seems to occur during the first four missions, as well as enhanced team situation awareness, teamwork, and taskwork knowledge. Team process behaviors should also transfer to the synthetic environment. These predictions will be tested in this study.

The following hypotheses are based on data previously collected from inexperienced participants and the assumption that our synthetic task environment captures aspects of the task in the field of practice relevant to team cognition. To test these predictions, the data obtained from experienced participants will be compared to data from inexperienced participants in Experiments 1 and 2 and in some cases, as appropriate, to data from participants in earlier CERTT UAV-STE experiments.

H3.1 Acquisition of team performance skill during the first four missions should occur at a faster rate for experienced operators, due to transfer of cognitive skill. For similar reasons, the drop that occurs for inexperienced operators with increases in workload, should not be as great for experienced operators.

H3.2 Team process behaviors (coordination, decision making, leadership, etc.) should be superior for experienced operators and should be revealed in process ratings and critical incident process scores.

H3.3 Measures of team situation awareness, taskwork knowledge, and teamwork knowledge should reflect differences, especially early in the task, between experienced and inexperienced participants.

4.10 Experiment 3: Method

4.10.1 Participants

Ideally we would have preferred to use intact teams of three Predator UAV operators for this experiment. If the UAV-STE *is* indeed faithful to the cognitive and team aspects of the operational task, then these operational teams would provide the best performance benchmark. However, subject matter experts are difficult to obtain, even in times of peace. In 2003 this problem was exacerbated due to the war with Iraq. We therefore decided to collect data from intact teams who had experience interacting in similar ways (i.e., ideally over headsets and through computers in a command-and-control like task). Further, we decide to test expert teams who varied somewhat in their mode of interaction. We felt that this would provide us with a way of distinguishing those factors that lead to expertise on this task.

Five three-person teams voluntarily participated in one 7-hour session. Team members had previously worked together as a team in various settings with most having worked on aviation, command-and-control, or military tasks. We recruited participants in the ASU East area who had worked together on other tasks because we believed that participants who were familiar with fellow team members would have better teamwork knowledge, which could affect team cognition and performance. A brief description of each team follows:

Team 1: Flight Instructor Team. Team 1 was made up of three male flight instructors. Two were commercial pilots while the other member was an airline transport pilot. The three had flown together in the same aircraft in combinations of two. When flying together, they trained each other in safety procedures. For the most part, these three instructors worked together to train flight students. Outside of the cockpit, the three men interacted together for more than a year. They reported that they all knew each other moderately well or very well.

Team 2: Video Game Team. Team 2 consisted of three males who were experienced in playing an on-line video game together. The game, Counter-Strike, provides the player with a simulation of what a trained counter-terrorist unit or terrorist unit experiences. It is a team-based game featuring one team playing the role of the terrorist and the other team members playing the role of the counter-terrorists. The three men reported that they played this on-line video game for several hours (2-3) on a daily basis for more than one year. The game is played in a distributed fashion using headsets to communicate over the internet. All team members considered themselves to be experts at this game and they all reported knowing each other very well.

Team 3: UAV Design Team. Team 3 was made up of three males who worked together at a company that designed and built UAVs. Two of the men worked together with one another more so than they did with the third person who was relatively new at the company and who worked on separate components of the engineering projects. The two men who were more familiar with one another interacted during test flights of the UAVs. One of them was the test pilot and the other was the navigator or operator. The team had worked together for less than one year but their interaction was fairly frequent. The two men who were more familiar with one another reported that they knew each other very well and reported knowing the newer co-worker moderately well.

Team 4: Flight Student Team. Team 4 was composed of three male college students enrolled in a flight training program at ASU East. On three to four occasions, these three flight students had flown in formation together where two of them each piloted a plane and the third person was a passenger in one of the planes. However, they reported that they all three worked together, communicating over headsets, to coordinate the flight. These three students have also interacted on several group projects in the classroom. Overall, the three reported knowing each other moderately well.

Team 5: CERTT Experimenters Team. Team 5 was made up of three male Ph.D. students who work in the CERTT Lab. The three students interact very closely as part of a small research team. They also have extensive knowledge and experience in the CERTT Lab in the specific areas of (1) conducting experiments in the UAV-STE, (2) designing measures, and (3) analyzing data. The three experimenters have worked together for several years (less than 5) on a daily basis. In addition, they have interacted occasionally for leisure purposes and report knowing each other either moderately or very well.

Participants were compensated by the payment of $10.00 per person hour with each of the three team-members on the highest-performing (non-experimenter) team receiving a $100.00 bonus. The participants were randomly assigned to teams and to roles (AVO, PLO, or DEMPC), with the exception of Team 3, the UAV design team. Two of these team members were experienced in flying and navigating UAVs; therefore to capitalize on this expertise, we assigned them to the AVO and DEMPC roles, respectively. The third team member who did not have experience operating UAVs was assigned to the PLO role.

4.10.2 Equipment and Materials

The experiment took place in the CERTT Lab at ASU East. This lab had been configured for the UAV-STE described previously. Equipment and materials for this experiment were identical to those described under Experiment 2 of this report.

4.10.3 Measures

The measures that were used in this experiment were identical to those described earlier (see Experiment 2) with two exceptions. First, we administered a paper-based demographics debriefing questionnaire with seven questions (see Appendix O). Second, a debriefing measure, which is also provided in Appendix P, was added to assess prior familiarity with the other members of the team.

4.10.4 Procedure

The procedure for this experiment was identical to that of Experiment 2 except that all five teams were run in the co-located condition.

4.11 Experiment 3: Results

4.11.1 Team Performance

Table 65 shows the mean performance for the five expert teams at each mission. Figure 30 shows the performance of each of the five teams in this experiment compared to the average performance of all teams of each of the four preceding experiments conducted in CERTT Lab's UAV-STE. (For naming convention we refer to the four previous studies done in the CERTT UAV-STE as AF1, AF2, AF3, and AF4 and the current one as AF5. These include Experiments 1, 2, and 3 of this project. AF3 maps onto Experiment 1 of this project, AF4 maps onto Experiment 2, and AF5 to Experiment 3). As can be seen, not only did two of the expert teams fail to outperform other non-expert teams, but the UAV Design Team scored drastically lower than other teams during Mission 4. Three of the expert teams, however, seem to have outperformed previous teams. These include the Experimenter, the Flight Student, and the Video Game Team.

Table 65
Descriptive Statistics for Team Performance at Each Mission

Mission	Mean	Standard Deviation	Minimum	Maximum
1	341.08	142.46	237.90	586.82
2	442.96	118.98	302.93	583.77
3	508.85	83.02	399.46	616.64
4	472.20	159.78	223.54	608.68
5	389.65	85.52	287.89	497.65

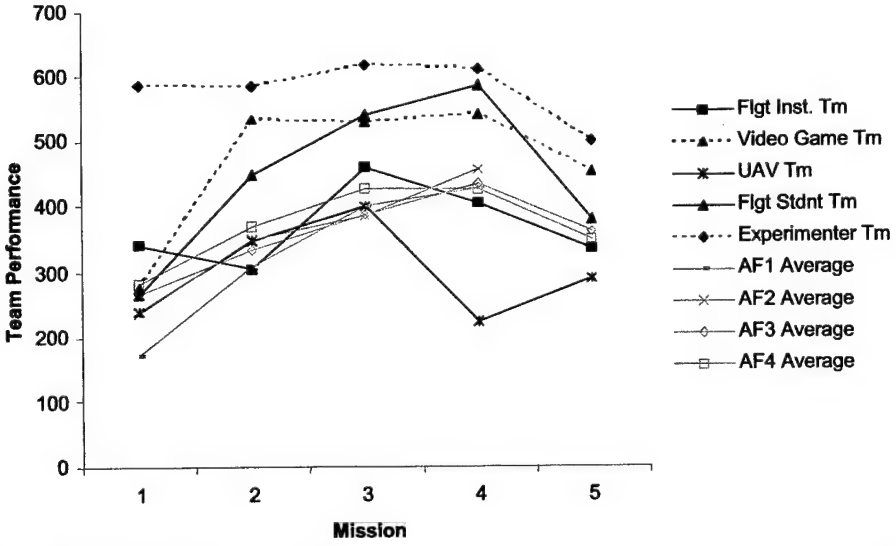


Figure 30. Each expert team's performance and the average of other teams' performance in previous experiments at the first five missions.

The performance of each expert team was compared to the mean performance of all other teams (a total of 69 teams) from the four other experiments. For Missions 1-4, expert teams' performance was compared to the mean of all 69 teams; however, for Mission 5, the expert teams' performance was only compared to the performance of the 40 teams in AF3 and AF4, as these were the only other experiments that included a workload manipulation. Table 66 indicates those expert teams that performed 1.5 standard deviations above (+) or below (-) the mean performance of the other teams from the first four experiments.

Table 66
Expert Teams who Achieved Performance Scores Ranging Outside |1.5| Standard Deviations of Non-Expert Teams' Performance at Each Mission

	Flght Inst. Team	Video Game Team	UAV Team	Flght Student Team	Experimenter Team
Mission 1					+
Mission 2		+			+
Mission 3		+		+	+
Mission 4			-	+	+
Mission 5		+			+

Again, the Experimenter, Flight Student, and Video Game teams excelled above the other expert teams and above the mean performance found in previous experiments. Looking again at the graph in Figure 30 it is also clear that these three expert teams also acquired team skill more quickly than the average teams. That is, it takes average teams about four missions working together as a team following individual training to reach asymptotic levels of team performance. The three expert teams seemed to reach asymptotic levels of performance sooner—in three or fewer missions. The three expert teams also obtained a level of asymptotic performance significantly higher than that of other teams. In addition, with the exception of the Flight Student Team, they maintained superiority even in the face of a high workload mission. Thus rapid team learning, high asymptotic levels of performance, and effectiveness under high workload seem to be the hallmark of expert teams in this task and the findings specific to the three highest scoring expert teams supports Hypothesis 3.1.

4.11.2 Team Process

Critical incident process. Table 67 gives the descriptive critical incident process statistics at each mission across the expert teams in this experiment. Interestingly mean CIP decreases in later missions (but only for some teams; see Figure 31). Mission 5 was high workload, thus it is possible that process suffered as a result of increased task requirements, however Mission 4 was a typical scenario and presumably should be similar, in terms of critical incident process, to Missions 1-3. However, we speculate that expert teams may be acting like distributed teams in the previous studies by increasing the efficiency of their process behavior by omitting some nonessential behaviors from their repertoire.

Table 67
Descriptive Statistics for AF5 Critical Incident Process at Each Mission

Mission	Mean	Standard Deviation	Minimum	Maximum
1	.75	.16	.5	.9
2	.76	06	.7	.8
3	.74	.07	.7	.85
4	.69	.14	.45	.8
5	.66	.18	.44	.89

The variability across the five teams is highest in Missions 1, 4, and 5. These missions might thus be the most informative concerning differences among the expert team's critical incident process scores. Referring to Figure 31, AF5 critical incident process scores were in most cases within or above the mission averages from earlier experiments. It is also interesting to note that four out of five expert teams *started* off with higher than normal critical incident process. This is probably one of the benefits of having worked together beforehand, plus transfer of team cognition, which also corresponds to the rapid acquisition of skilled team performance.

Next, looking at the expert team's critical incident process singly, it is clear that the Video Game Team and the Experimenter Team were the highest (Figure 31) across all missions. Interestingly, the Flight Instructor Team had drastic reductions in critical incident process scores at Missions 4 and 5. This team presumably abandoned regimented patterns for coordinating early on. Perhaps they misperceived the task and tried to rely on other forms of coordination. Similar arguments can be made with respect to the UAV Team, who started out high on critical incident process but apparently abandoned the standard type of coordination after Mission 1, and then again after Mission 4. On the other hand, the Video Game Team and the Experimenter Team generally had high critical incident process and appropriately enough these were the two highest performing teams.

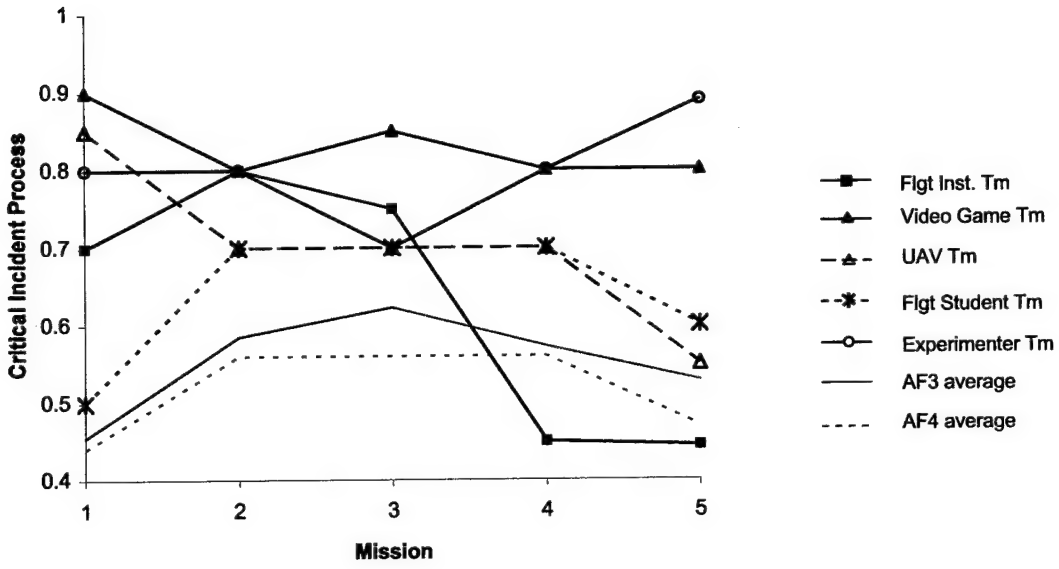


Figure 31. AF5 critical incident process across missions for each team with average critical incident process across first 5 missions for AF 3 and AF 4.

We examined how unusually high or low expert critical incident process scores were relative to other teams' critical incident process scores (note that comparisons are only against AF3 and AF4 teams; the critical incident process scoring instrument changed after AF2) at each mission. Table 68 presents unusually high or low critical incident process scores for each team's missions in Experiment 5. In terms of standard deviations away from standard mission critical incident process scores, Table 68 points out the most unusually high ($\geq$ 1.5 SD; +) or low ($\leq$ -1.5 SD; –) critical incident process scores for each mission-at-team in Experiment 3. There were no (–) marks indicating that experts never obtained below expectation critical incident process scores. On the contrary, the only marks were (+), not surprising given that these teams had had experience working together. Overall the Video Game Team and the Experimenter Team had the most (+) marks. These correspond to the highest performing benchmark teams suggesting that good critical incident process may underlie the highest benchmark performers.

Table 68
Mission-at-team Experiment 5 Critical Incident Process Z-scores Ranging Higher than |1.5| Standard Deviation at Each Mission

	Flight Inst Team	Video Game Team	UAV Team	Flight Student Team	Experimenter Team
Mission 1	+	+	+		+
Mission 2	+	+			+
Mission 3		+			
Mission 4		+			+
Mission 5		+			+

Summary process. The component process items were averaged for an overall summary process score. Table 69 lists the descriptive statistics for summary process across teams in AF5. The range of the summary process score is from 1 to 5. In general, the expert teams' average summary process was high in all missions, with the highest averages obtaining for Missions 3 and 4 and a drop off at Mission 5, the high workload mission. Again, the Mission 5 drop off held for only a couple of expert teams (Figure 32).

Table 69
Descriptive Statistics for AF5 Summary Process at Each Mission

Mission	Mean	Standard Deviation	Minimum	Maximum
1	4.15	.49	3.63	4.88
2	4.03	.92	3.00	5
3	4.53	.46	3.88	5
4	4.58	.56	3.75	5
5	4.18	.90	2.88	5

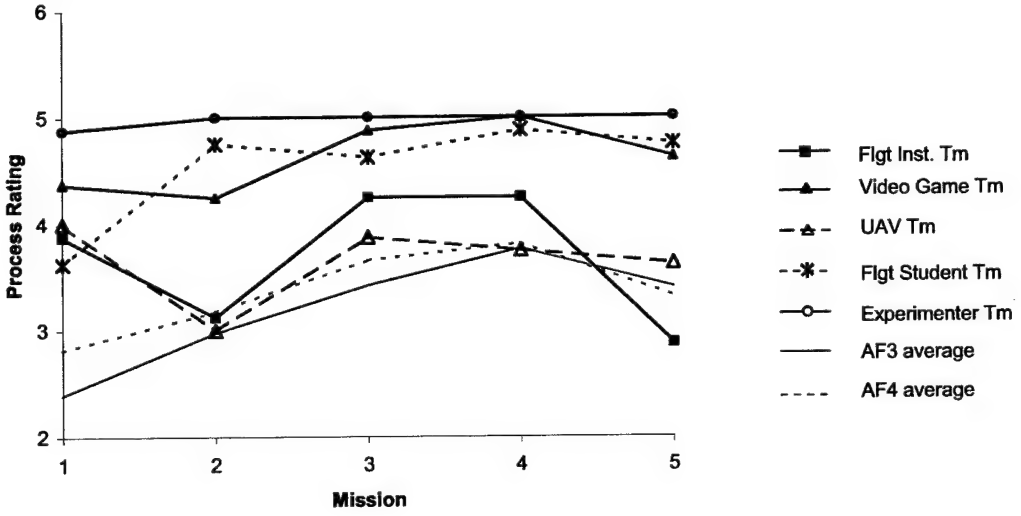


Figure 32. Experiment AF5 summary process across missions for each team with average summary process scores across first 5 missions for AF3 and AF4.

In comparison to the other experiments in which summary process was measured (AF3 & AF4), expert summary process scores tended to be high (Figure 32). However some expert teams were higher than others on this measure. The expert teams with the highest summary process scores were the Video Game Team, the Experimenters Team, and the Flight Student Team. The teams with the lowest summary process scores were the Flight Instructor Team and the UAV Design Team. Although these teams had the lowest summary process scores relative to the other expert teams, with a few exceptions they were above the average summary process of AF3 and AF4 teams at each mission. The summary process findings are in agreement with the team performance and critical incident process findings described above.

Turning to the question of how unusually high or low are the expert summary process scores relative to all other summary process scores, Table 70 lists the missions in which each team was either 1.5 SD higher (+) or lower (-) than expected. No cell in Table 70 contains a (–), so expert summary process scores were never 1.5 SD lower than expectations. On the other hand, quite a few expert missions showed higher than expected summary process scores. Most striking, the top scoring Video Game and Experimenter Teams obtained unusually high process summary scores in five out of five missions. These high-process teams were followed by the Flight Student Team, the UAV Design Team, and the Flight Instructor Team, in that order. These results map onto the performance rankings of these teams indicating that the summary process ratings do show validity in terms of team performance.

Table 70
Mission-at-team AF5 Summary Process Z-scores Ranging Higher than |1.5|Standard Deviations at Each Mission

	Flight Instructor Team	Video Game Team	UAV Team	Flight Student Team	Experimenter Team
Mission 1		+	+		+
Mission 2		+		+	+
Mission 3		+		+	+
Mission 4		+			+
Mission 5		+		+	+

In general the three high-scoring expert teams also exhibited superior team process behaviors. These behaviors were often seen in early missions as well as in the high workload mission. These results are parallel to those of performance and generally support the second hypothesis (3.2) regarding superior process for the expert teams.

4.11.3 Situation Awareness

Table 71 shows descriptive statistics for situation awareness accuracy on repeated and non-repeated queries on a mission-by-mission basis. A single data point was missing for the repeated query at Mission 1 (Team 1).

Figure 33 shows the situation awareness accuracy to the repeated query for each team in AF5 compared to the average situation awareness of all teams in each of the four preceding experiments conducted in the UAV-STE. Only situation awareness data from Missions 1 through 4 were used for AF1 and AF2, as these experiments did not involve the workload manipulation during Mission 5. As can be seen, two of the expert teams maintained levels of situation awareness accuracy similar to the average for each of the other four experiments. However, the Video Game Team, the Experimenter Team, and the Flight Student Team achieved perfect accuracy scores during at least one mission. Not surprisingly, the Experimenter Team members, who were very familiar with the task and the situation awareness queries, achieved perfect accuracy on the repeated query in all low workload missions, supporting our previous claim that teams learn the number of targets they can acquire (the maximum; a ceiling effect). The experimenters knew that there were nine targets in the low workload missions and that they

were able to acquire them all. Other teams less familiar with the situation took more trials to learn this.

Table 71
Descriptive Statistics for Situation Awareness at each Mission (N = 5)

Mission	Mean	Standard Deviation	Minimum	Maximum
1*	1.00	1.41	0.00	3.00
2	.80	1.30	0.00	3.00
3	1.60	1.14	0.00	3.00
4	2.60	.54	2.00	3.00
5	.80	1.30	.00	3.00
1	2.20	.84	1.00	3.00
2	2.20	.84	1.00	3.00
3	1.80	.84	1.00	3.00
4	2.00	.71	1.00	3.00
5	2.00	1.00	1.00	3.00

* N = 4

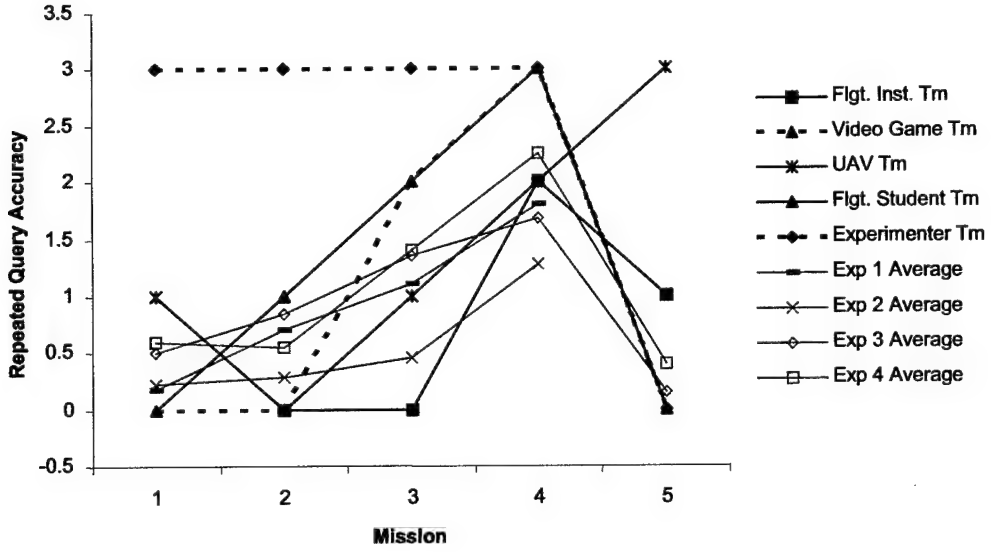


Figure 33. Situation awareness accuracy on the repeated query for each expert team and for all teams in AF1 through AF4 at each mission.

Figure 34 shows the average of each expert team's and the average of all AF3 and AF4 teams' accuracy to the non-repeated queries. Non-repeated queries were not asked in Experiments 1 and 2. Due to the nature of the non-repeated queries, Figure 34 does not take mission into consideration and instead presents an average of each team's accuracy to the queries across missions. Again, the Experimenter Team's average accuracy was nearly perfect. All other expert teams achieved levels of situation awareness accuracy to the non-repeated queries similar to non-expert teams in AF3 and AF4. The intricacies of the Experimenter Team's knowledge are apparent here in terms of expectations and background knowledge.

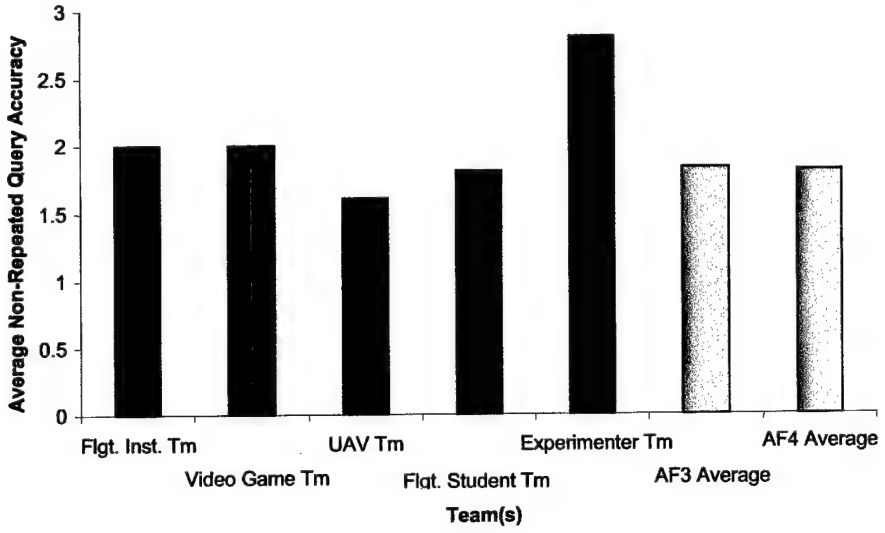


Figure 34. Average situation awareness on the non-repeated query for each expert team and for all teams in AF3 and AF4.

Each expert team's situation awareness accuracy on the repeated query was compared to the average of all other non-expert teams' accuracy to the repeated query. For Missions 1-4, expert teams' situation awareness was compared to the situation awareness of all other 69 teams; however, for Mission 5, the expert teams' situation awareness was only compared to the situation awareness of the 40 teams in AF3 and AF4, as these were the only other experiments that included a high workload condition.

Table 72 indicates those expert teams who achieved 1.5 standard deviations above (+) the average situation awareness of the other teams from the first four experiments. Due to floor effects, it was not possible for expert teams to score below 1.5 standard deviations of the non-expert teams' scores, as many non-expert teams scored the lowest possible. Furthermore, due to a ceiling effect at Mission 4, it was also not possible for expert teams to score 1.5 standard deviations above the mean of the non-experts' situation awareness accuracy during this mission, as many of the non-expert teams achieved extreme scores (i.e., either the lowest or highest possible score).

Table 72
Teams who Achieved Situation Awareness Accuracy Scores on the Repeated Query Ranging Above |1.5| Standard Deviations of Non-Expert Teams at each Mission

	Flight Instructor Team	Video Game Team	UAV Team	Flight Student Team	Experimenter Team
Mission 1					+
Mission 2					+
Mission 3					+
Mission 4					
Mission 5			+		

Results indicate that the Experimenter Team and the UAV Design team were significantly superior to the other teams in terms of accuracy on the repeated query. The fact that the Experimenter Team did so well on this measure can be attributed to this team's extensive knowledge of the experimental situation including the nature of the upcoming situation awareness queries (i.e., expertise). Although on the surface, this result supports our third hypothesis (H3.3) we are not convinced that the measure truly reflects what we mean by team situation awareness.

4.11.4 Taskwork Knowledge

The means and standard deviations as well as the minimum and maximum scores for overall taskwork accuracy during the knowledge session for AF5 teams can be seen in Table 73. Figure 35 shows the average taskwork knowledge scores for each experiment conducted in the UAV-STE as well as the taskwork knowledge of the 5 expert teams. Knowledge scores for AF1, AF2, and AF3 were each based on the second knowledge session from those experiments.

Table 73
Taskwork Knowledge Scores for Experiment 3 Teams

	Mean	Standard Deviation	Minimum	Maximum
Overall Accuracy	.53	.08	.44	.64
Positional Knowledge	.00	. 82	-.94	1.24
Interpositional Knowledge	.00	.85	-.73	1.09
Intrateam Similarity	.41	.12	.29	.56
Holistic Accuracy	.56	.12	.39	.73

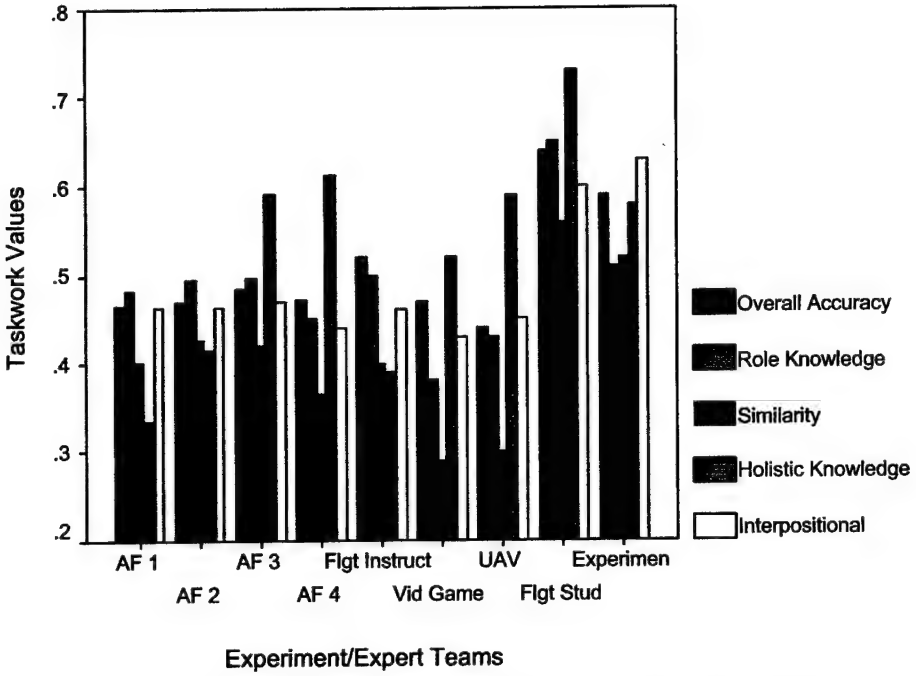


Figure 35. Average taskwork values for all experiments and expert teams.

The team comprised of Flight Students excelled in all forms of taskwork knowledge. Interestingly, the Flight Students knew more about taskwork than even the Experimenter Team, who nonetheless possessed better-than average overall and interpositional taskwork knowledge. Interestingly, the high-scoring video game team had significantly lower role knowledge scores than teams on average.

Table 74 indicates those expert teams that performed 1.5 standard deviations above (+) or below (-) the mean performance of the other teams from the first four experiments.

Table 74
Indications of Expert Teams who Achieved Taskwork Knowledge Scores Above those of all other Non-Expert Teams

	Flight Inst. Team	Video Game Team	UAV Team	Flight Student Team	Experimenter Team
Overall Accuracy				+	+
Positional Knowledge		-		+	
Interpositional Knowledge				+	+
Intrateam Similarity				+	
Holistic Accuracy				+	

In general, the two highest-performing expert teams, the Video Game team and the Experimenter Team do not seem to excel above non-expert teams in taskwork knowledge. However, the Flight Student Team, the third highest-performing expert team seemed to have exceptional levels of taskwork knowledge. Perhaps their status as students made the declarative knowledge acquisition somewhat natural. Overall the hypothesis (H3.3) regarding superior taskwork knowledge for expert teams is not strongly supported by these results.

4.11.5 Teamwork Knowledge

Only measures of teamwork scores from Experiments AF 3 and AF4 are used from comparison in this analysis because these scores were obtained using the same instrument as Experiment 5. Table 75 shows descriptive statistics for overall teamwork accuracy scores obtained from all teams who participated in Experiments AF3, AF4, and AF5. There are virtually no differences among the three mean overall teamwork knowledge scores.

Table 75
Descriptive Statistics for Teamwork Knowledge Scores

Experiment	N	Mean	Standard Deviation	Minimum	Maximum
AF3	20	24.82	1.47	22.67	28.00
AF4	20	23.18	2.73	17.00	28.67
AF5	5	23.47	3.58	18.00	26.67

Table 76 allows us to compare teamwork knowledge accuracy scores obtained from the five expert teams who participated in Experiment AF5.

Table76
Descriptive Statistics for each Team's Teamwork Knowledge Accuracy Scores at Experiment AF5.

Team	Mean	Standard Deviation	Minimum	Maximum
1. Flight Instr. Tm	18	5.21	15	24
2. Video Game	22	4.36	19	27
3. UAV Tm.	24.33	1.53	23	26
4. Flight Student	26.67	3.56	24	30
5. Experimenter Tm	26.33	1.53	25	28

The Flight Instructor team and the Video Game team had the lowest mean teamwork scores and were the only teams to deviate more than 1.5 standard deviations (in a negative direction) from the mean teamwork accuracy score for AF3 and AF4 teams. Although the Flight Instructor Team was not one of the high-performing expert teams, the Video Game Team was, supporting other findings in which this declarative knowledge of teamwork correlates little with team performance. Again there is little support for our third hypothesis that concerns teamwork knowledge (H3.3).

4.12. Experiment 3: Discussion

Out of the five rather different types of expert teams, there were three teams that achieved performance levels on the UAV-STE that were clearly superior to the norm based on all previous UAV-STE teams. These teams were the Video Game Team, the Flight Student Team, and the Experimenter Team. One could argue that compared to the Flight Instructor Team and the UAV Design Team, these three teams had prior experience most resembling the command-and-control type of task that characterizes the UAV-STE. These three groups also seemed to be well versed at communicating using headsets. On the other hand, the Flight Instructor Team worked together as teachers and the UAV Design Team worked together on a design team. Although two members of the UAV Design Team have experience operating UAVs, their previous experience appeared to interfere with how the UAV synthetic task is performed. The UAV-STE task was deceptively similar (on the surface) to the task familiar to the UAV Design Team. In fact, the UAV design team reported experiencing difficulty in adjusting to these differences. Therefore, perhaps team member familiarity in the context of an isomorphic team task is what is most important for transfer of team cognition.

The teams that excelled demonstrated a faster-than-normal rate of team skill acquisition with the Experimenter Team achieving asymptotic performance in the first mission. Two of thee teams also excelled under high workload (Experimenter and Video Game). Taken together these findings support our hypothesis that acquisition of team performance skill during the first four missions should occur at a faster rate for experienced operators, due to transfer of team cognitive skill and the drop that occurs for inexperienced operators with increases in workload, should not be as great for experienced operators. Also asymptotic performance of the three top expert teams was generally about 200 points better than typical UAV-STE teams.

Our second hypothesis stated that team process behaviors (coordination, decision making, leadership, etc.) should be superior for experienced operators and should be revealed in process ratings and critical incident process scores. This hypothesis is also supported. The three top expert teams tended to exhibit better team process behavior from the start, supporting the claim that the coordination or interaction behavior analogous to individual cognitive processing is what is critical for good performance on this task and what enables the top expert teams to excel.

Our third hypothesis that measures of team situation awareness, taskwork knowledge, and teamwork knowledge should reflect differences especially early in the task between experienced and inexperienced participants was not supported. Although the Experimenter Team demonstrated high levels of situation awareness from the start, we suspect that this is for the wrong reasons. Results pertaining to taskwork and teamwork knowledge did not suggest that the expert teams or even top expert teams were generally superior on these measures. The one exception was the Flight Student Team who excelled at taskwork knowledge. However, this may have more to do with the fact that this was a group of students accustomed to acquiring large amounts of declarative knowledge, than the fact that they were an expert team.

In conclusion, teams who were expert at command-and-control tasks did very well on the UAV-STE. These top teams also had very good team process behaviors, which may explain the superior performance and rapid rate of skill acquisition as a team. Whereas taskwork and

teamwork knowledge may be a prerequisite for acceptable performance, it seemed to be team process behaviors that best distinguished top teams from the other teams.

In the next two sections we summarize data across AF1, AF2, AF3, and AF4 that speak to the issue of the role of individual differences in team performance and cognition, and the reliability and validity of our knowledge measures.

4.13 Archival Analysis of Individual and Role-Associated Factors

This part of the project was designed to address the third objective which is to *investigate the relation between individual characteristics and team cognition and performance through an archival analysis on data from four previously conducted CERTT UAV-STE experiments.* This objective involved the following four tasks: Task 1: Assemble data collected from four CERTT-UAV studies, Task 2: Across the four studies, attempt to identify individual and team differences (cognitive and otherwise) that account for significant variance in team performance, Task 3: Explore the impact of individual differences associated with team role on team performance and cognition, and
Task 4: Explore the use of voice stress as an index of individual arousal during mission performance. The first task is described in the methods section and the other three tasks are described in the results section.

This objective was motivated by some preliminary findings in Experiment 1, which indicated a possible relation between team performance and the working memory capacity of individual team members. In addition this relationship seemed to depend on the role of the team member, with the DEMPC seeming to weigh more heavily with regard to this relationship than other team members. As we pointed out in the introduction of this report, relatively little is known about the impact of individual differences on team performance and cognition. Although there are some data indicating a relationship between cognitive abilities and group performance, there is virtually nothing on cognitive abilities in heterogeneous teams in which each individual has a specific role on the team. It is interesting to ask whether the impact of the individual characteristic depends on the role that an individual assumes on the team. That is, working memory capacity may be critical for some roles more so than others. Our UAV-STE provides a perfect setting for investigating the relation between individual differences, team role, and team performance.

There are also pragmatic motives driving this objective. By considering such differences among individuals, team roles, and between teams composed of different individuals (i.e., team differences due to team composition), we should be able to account for variance in team and individual performance that would otherwise be left unexplained. By identifying and removing this variance, this approach may allow us to detect more subtle effects of manipulations on team performance. Additionally, identifying individual differences that are critical for team performance is a necessary step toward efforts to improve team performance through team composition, focused training, or design aids.

Thus, toward our third objective, in Experiment 2 we collected additional data on working memory, and also some measures of verbal processing speed and voice data. Further, because statistical tests of individual and team differences are more powerful with larger samples, we

decided to look at these factors across all of the teams and individuals in Experiments 1 and 2 as well as in two experiments conducted previously under a separate effort. Total, there are 69 teams across these four experiments. As we did for the benchmarking experiment, we will again follow the convention of referring to the four experiments as AF1, AF2, AF3, and AF4 where AF3 corresponds to Experiment 1 conducted under this effort and AF4 corresponds to Experiment 2.

4.14 Archival Analysis of Individual and Role-Associated Factors: Methods

4.14.1 Participants

The 207 participants (69 teams) for these analyses came from four studies conducted in the CERTT lab (referred to as AF1, AF2, AF3, and AF4). AF3 and AF4 were presented earlier in this technical report as Experiments 1 and 2 whereas the other two experiments were part of an earlier three-year research effort funded by AFOSR (Cooke, et. al., 2001). The teams from these four studies were composed of student volunteers from NMSU. The characteristics of the samples are presented in the last two rows of Table 77. For all of the studies except AF4, the participant's organization was compensated rather than the participant.

Table 77
Procedural Characteristics of Four Air Force Studies

	AF1	AF2	AF3	AF4
# Missions	10	5	7	5
Workload	Constant	Constant	Missions 1-4 =LW Missions 5-7 =HW	Missions 1-4 =LW Missions 5-7 =HW
# Knowledge Sessions	4	3	2	1
Placement of Knowledge Session	1-after Mission 1 2-after Mission 4 3-after Mission 7 4-after Mission 9	1-after training 2-after mission 2 3-after all missions	1-after training 2-after all missions	1-after Mission 3
Mission Time	40 min	40 min	40 min	40 min
# Teams	11	18	20	20
# Sessions	3	2	2	1
Manipulations	None (acquisition task)	Knowledge (shared vs. none)	Dispersion Workload	Dispersion Workload
Participants	AF ROTC cadets	AF ROTC cadets	Campus organizations	Male students
Compensation	$6/hr to organization; $50 bonus to best team	$6/hr to organization; $50 bonus to best team	$6/hr to organization; $50 bonus to best team	$6/hr to individual; $50 bonus to best team

4.14.2 Equipment and Materials

All experiments were conducted in the CERTT Lab using the same equipment, although configured differently for AF3 and AF4's co-located vs. distributed experiments. Measures were for the most part the same as those described for Experiment 1. Differences will be described in the appropriate results sections. Some additional equipment was used to analyze the voice data. Audio data from the headsets were recorded in all experiments on an Alesis digital recorder. The output from the Alesis recorder was sent to a Terratec EWS88D ADAT SPDIF soundcard that used Samplitude Version 5.55 mastering and multitracking software to make digitized waveform files. Frequency data from these files were then analyzed using Time Frequency Representation software that was manufactured by Avaaz Innovations.

4.14.3 Measures

The measures that will be analyzed are those collected at an individual level and thus team process measures are not included here. Measures included in this analysis are listed in Table 78. Measures of team performance and team cognition (situation awareness, taskwork knowledge, and teamwork knowledge) were collected at both the individual and team levels. Measures that were only collected at the individual level include verbal working memory capacity, verbal processing speed, voice stress, NASA TLX, grade point average, and various demographic variables.

The variables in Table 78 were discussed in the Experiment 1 section except for the voice analysis data. In order to assess voice stress, we used the audio recordings that were made as participants performed the UAV task. Fundamental frequency, amplitude, and other voice parameters were recorded for each participant during particular missions. Comparisons were made between Missions 1 and 4 and Missions 4 and 5 to look for changes in voice stress over task acquisition and increasing workload, respectively.

Table 78
Measures Collected at the Individual Level Across Four UAV-STE Experiments

Measure	UAV Experiment			
	AF1	AF2	AF3	AF4
Performance	X	X	X	X
Situation Awareness	X	X	X	X
Taskwork Knowledge	X	X	X	X
Teamwork Knowledge	X	X	X	X
Verbal Working Memory			X	X
Verbal Processing Speed				X
Voice Stress Data	X	X	X	X
NASA TLX			X	X
GPA	X	X	X	X
Demographics	X	X	X	X

4.14.4 Procedure

The procedure described earlier was similar for each experiment (i.e., PowerPoint training, skills training, multiple missions) with variations occurring in mission environment, number of missions, and workload. Procedural differences between the studies are listed in Table 76. Measures of performance, voice, demographics (including GPA), process, and cognition were taken as stated previously.

4.15 Archival Analysis of Individual and Role-Associated Factors: Results

Our analyses were conducted to address the following questions: (1) How does variation on the individual characteristic relate to team performance, (2) How does the variation on the individual characteristic relate to the team-level of that same characteristic (in cases in which a team level measure exists), (3) How does variation on the individual characteristic specific to a team role relate to team performance, (4) How does variation on the individual characteristic specific with a team role relate to the team-level of that same characteristic (in cases in which a team level measure exists), and (5) How does variation on the individual characteristic specific to role relate to role-specific performance.

In most cases, in order to address the first and second questions, regression analyses were conducted in which the maximum and range of the individual scores on a team served as predictors for either the team performance criterion or the team-level measure of that same factor. A significant maximum indicates an importance of an individual score in terms of a team score. A significant range indicates the importance of variability among the individual scores to the team score. A significant interaction between the two means how important the maximum is in terms of the team score depends on how tightly or loosely dispersed the individual scores are. In order to test the role effects associated with Questions 3 and 4 a univariate analysis of covariance (ANCOVA) was run in which role and experiment served as independent variables. In order to test the role-specific effects of a variable on role-specific performance a multivariate analysis of variance (MANOVA) was run in which role and experiment served as independent variables and role-specific performance as dependent variables.

4.15.1 Individual Performance

This archival data analysis was performed in order to answer two of our research questions: (1) How does variation on individual performance relate to team performance, and (3) How does variation on individual performance interact with role to affect team performance. Questions 2 and 4 did not apply here since the team-level variable was the same as team performance. Question 5 did not apply because the role-specific variable was the same as role-specific performance.

Because Mission 4 was the point at which individuals and teams reached asymptotic levels of performance, Mission 4 performance was used as an estimate of individual and team performance across all four experiments. Moreover, because our individual performance scores were based on different task components they were not on the same scale and thus, individual performance scores were standardized before they were entered into this analysis.

Individual performance. A regression analysis was performed in order to address the first main research question, namely, how individual performance relates to team performance. The maximum performance score for each team as well as the range for performance scores for each team were used as independent variables. Results are presented in Table 79.

Table 79
Results from the Regression Analysis of Individual Performance

Source	*SS*	*df*	*MS*	*F*	*p*
Max	52, 119	1	52, 119	12.44	<.01*
Range	4, 350	1	4, 350	1.04	.31
Max*Range	40, 034	1	40, 034	9.56	<.01*
Total	434, 913	66			

* $p < .10$

There was a significant main effect of the maximum, with the occurrence of a high individual score predicting a high team score. There was also a significant interaction effect between maximum and range. Thus individual maximum predicts team performance best when all three individuals have high scores.

Role-specific performance. To determine the impact of each role's performance (that is AVO, PLO and DEMPC) on team performance, a univariate analyses of covariance (ANCOVA) was run:

> Team Performance = Experiment, AVO_Perf, PLO_ Perf, DEMPC_ Perf,
> Experiment*AVO_ Perf, Experiment*PLO_ Perf, Experiment*DEMPC_ Perf

Again, performance data from Mission 4 were used as a summary variable for this ANCOVA. From the interaction terms in the model, heterogeneity of slopes was evaluated to identify differences in the role-team relationship across experiments. Looking at the results in Table 80 there are differences in the DEMPC performance-team performance relationship depending on the experiment. Since AVO and PLO relationships did not change depending on experiment, we conclude that AVOs and PLOs have an overall main effect on team performance; that is, having high performing AVOs and PLOs is associated with higher team performance.

Table 80
Results of the Univariate ANCOVA Examining the Relationship between Role Performance and Team Performance

Source	*SS*	*df*	*MS*	*F*
Exp	10.1	3	10.1	8.44***
AVO_Perf	16.1	1	16.1	40.03***
PLO_Perf	2.1	1	2.1	5.18**
DEMPC_Perf	1.8	1	1.8	4.51**
Exp*AVO	.1	3	.1	.06
Exp*PLO	2.5	3	2.5	2.12
Exp*DEM	2.9	3	2.9	2.41*

* $p < .10$ ** $p < .05$ *** $p < .01$

As mentioned, in testing for heterogeneity of the individual-team performance relationships, it became clear that this relationship for the DEMPC role was different based on the particular experiment. *Post hoc* regressions were run in order to determine the DEMPC-team performance relationship at each experiment (see Table 81). In particular, the DEMPC-team relationship was very strong and positive in the data collected for Experiments AF1 and AF4 and a bit weaker, yet still positive for Experiment AF3. The DEMPC-team relationship was weakest for Experiment AF2.

Table 81
Results of Post-hoc ANCOVA Examining Direction and Significance of Relationship Between DEMPC and Team Performance Across Experiments

DEMPC-Team relationship at each experiment	N	*t*	*p*	Pearson *r*
DEMPC*Team Perf at AF1	11	1.82	.10	.52*
DEMPC*Team Perf at AF2	18	.71	.48	.17
DEMPC*Team Perf at AF3	20	-1.39	.18	.32
DEMPC*Team Perf at AF4	20	2.36	.03	.50**

* $p < .10$.** $p < .05$

Summary. The analysis examining individual-team and role-team performance relationships revealed several significant results. Not surprisingly, high scoring teams were composed of high scoring individuals. And in most cases role did not matter for this factor. It is important to note here that team performance was measured at the holistic level and was not measured as a sum or average of individual performances.

Additionally, the magnitude of the F statistics suggests that the AVO role bore the strongest and most consistent relationship to team performance, followed by PLO and then by DEMPC. The question arises why DEMPC performance would not be as predictive of team performance? One possible explanation is that the individual DEMPC score is based largely on effective route planning which does not directly translate into good photos of targets. Overall, however, it would not appear that role was a critical factor in the relationship between individual and team performance.

4.15.2 Individual Situation Awareness

This section presents an archival analysis of situation awareness data collected over the four experiments conducted in the context of the UAV-STE. The purpose of these analyses was to systematically look for patterns in the relationship between individual situation awareness and team performance and situation awareness across the four experiments. These analyses addressed the five questions mentioned above but with regard to individual and team situation awareness and performance.

Because situation awareness was not measured in a manner that was consistent across all four experiments, some preliminary steps were taken to prepare the data for the archival analysis. First, in each experiment, a series of situation awareness queries were administered at randomly determined points in the mission. However, because the set of queries was not the same in each experiment, only those queries common to all four experiments were used in the current analyses. There was one query common to all experiments, the "repeated query," which asked the team to predict the number of targets their team would be able to photograph successfully by the end of their 40-minute mission. In each experiment, this query was repeated at every mission. Second, because the situation awareness data were not scored in a manner that was consistent across all four experiments, the data were re-scored such that each individual accuracy score was represented as either 0 (inaccurate) or 1 (accurate).

Third, various manipulations used in three of the experiments were not incorporated into these archival analyses. Although Experiments AF3 and AF4 involved the same manipulation (co-located vs. distributed), it was not incorporated into these archival analysis, as the manipulation failed to show an effect on situation awareness in both of those individual experiments. Finally, because Mission 4 was the point at which individuals and teams reached asymptotic levels of performance, Mission 4 performance will be used as an estimate of individual and team performance across the experiments. Likewise, situation awareness accuracy measured at Mission 4 will serve as the measurements of individual and team situation awareness.

Individual situation awareness. These analyses address the first two research questions listed above, namely, how does individual situation awareness relate to team performance and team situation awareness? For instance, is having one person on the team with high situation awareness enough to have a high performing team?

Because the situation awareness scores for individuals were 0 or 1 the maximum/range regression analysis did not apply. Instead, correlations were run to relate individual situation awareness (the sum of each individual's situation awareness accuracy) with team performance and also individual situation awareness with team situation awareness (a holistic measure of situation awareness in which team members reached consensus on the situation awareness query). Situation awareness data were missing for seven teams. A significant correlation between individual situation awareness and team performance was found, indicating that the more members on a team with accurate responses to the situation awareness query, the higher team performance, $r = .35$, $p < .01$, $n = 62$. For the second correlation, only data from Experiments AF3 and AF4 were used, as situation awareness was not measured at the team level during Experiments AF1 and AF2. Furthermore, there were missing data for two teams. A large, positive correlation was found, $r = .83$, $p < .01$, $n = 38$, suggesting that teams made up of individuals with high individual accuracy to the situation awareness query are more accurate on the team (holistic) situation awareness query.

These correlations suggest that individual situation awareness is an important contributor to team success. The ability that the individuals acquire over the course of the first three missions to accurately respond to the situation awareness query in Mission 4 plays a role in how those individuals come together to (1) perform as a team, and (2) have good team situation awareness.

Role-specific situation awareness. This section addresses the final three research questions, that is, how does the situation awareness associated with a particular role (i.e., AVO, PLO, and DEMPC) relate (1) to team performance, (2) to team situation awareness, and (3) to role-specific performance.

To determine the impact of each role's situation awareness on team performance, a univariate analyses of covariance (ANCOVA) was performed. As stated above, data from Mission 4 served as a summary of the participants' situation awareness accuracy for the duration of the experiment. Data were missing for seven cases, which resulted in a total of 62 cases for analysis. The following model was run:

> Team Performance = Experiment, AVO_SA, PLO_SA, DEMPC_SA,
> Experiment*AVO_SA, Experiment*PLO_SA, Experiment*DEMPC_SA

To test for the heterogeneity of slopes across the experiments, we first interpreted the interactions in the model. The F-values in Table 82 show that none of the interaction effects were significant, which implies that the relationship between each role's situation awareness and team performance was not significantly different across experiments. There was a significant experiment effect due to changes in mean levels of team performance (AF2>AF3>AF4>AF1). Of the three roles, only AVOs' situation awareness predicted team performance.

Table 82
Results of the Univariate ANCOVA Examining the Relationship Between Role Situation Awareness and Team Performance

Source	Num *df*	Den *df*	*F*
Exp	3	46	4.49**
AVO SA	1	46	2.80*
PLO SA	1	46	.00
DEMPC SA	1	46	2.25
Exp*AVO SA	3	46	.61
Exp*PLO SA	3	46	.03
Exp*DEMPC SA	3	46	.96

**$p<.01$. *$p\leq.10$

The next research question addresses the effect of each role's situation awareness on team, or holistic, situation awareness. Because situation awareness was only measured at the team level during Experiments AF3 and AF4, only data from those experiments will be used in the analysis. Again, a univariate ANCOVA was used:

> Team SA = Experiment, AVO_SA, PLO_SA, DEMPC_SA, Experiment*AVO_SA,
> Experiment*PLO_SA, Experiment*DEMPC_SA

The F-values in Table 83 show that the relationship between DEMPC situation awareness and team situation awareness was not homogeneous across experiments. Testing the simple effects, we found that in Experiment AF3, DEMPC situation awareness was positively related to team situation awareness, $t(16) = 3.11, p < .01, \beta = .61$. In Experiment AF4, the correlation between

DEMPC situation awareness and team situation was one, indicating that DEMPC situation awareness in AF4 perfectly predicted team situation awareness. Due to the interaction found for DEMPC, the main effects of role situation awareness on team situation awareness are not interpreted here. However, as the test for simple effects suggest, the DEMPCs' situation awareness bore the strongest relationship to holistic situation awareness.

Table 83
Results of the Univariate ANCOVA Examining the Relationship Between Role Situation Awareness and Team Situation Awareness

Source	Num *df*	Den *df*	*F*
Exp	1	30	.34
AVO SA	1	30	.32
PLO SA	1	30	2.65
DEMPC SA	1	30	34.48**
Exp*AVO SA	1	30	.32
Exp*PLO SA	1	30	2.65
Exp*DEMPC SA	1	30	5.18*

*$p<.05$. **$p<.01$

The analyses on the effects of role demonstrate that the DEMPC's situation awareness is critical to the team's situation awareness. Recall that the DEMPC is the coordinator of the mission. The DEMPC has a global view of the mission at all times whereas the other roles have only a snapshot view of the mission at any one point in time. For this reason, it is not surprising that the DEMPC's awareness of how many targets the team has visited and how many are left to be photographed is so critical to the accuracy of the holistic situation awareness query, which asks the team to reach consensus on how many targets they will be able to successfully photograph by the end of the mission.

The impact of each role's situation awareness on performance associated with each role was assessed using a multivariate analysis of variance (MANOVA). The following model was run:

> AVO Performance PLO Performance DEMPC Performance = Experiment, AVO_SA, PLO_SA, DEMPC_SA, Experiment*AVO_SA, Experiment*PLO_SA, Experiment*DEMPC_SA

In testing for heterogeneity of the role situation awareness-role performance relationships, it became clear that the relationship between DEMPCs' situation awareness and role performance differed across experiments (see Table 84 for statistics). *Post-hoc* comparisons were run to isolate which individual role performance was finding an interaction effect between DEMPC situation awareness and experiment. The partial relationships for the interaction with AVO performance and DEMPC performance were not significant, $F(3, 46) < 1$ and $F(3, 46) = 1.07$, respectively. However, this effect on PLO performance did marginally differ across experiments, $F(3, 46) = 2.02$, $p = .12$. Looking at simple effects of DEMPC situation awareness at each experiment, we determined that DEMPC situation awareness was positively related to PLO performance for Experiment AF3, $\beta = .30$, $t(57) = 2.22$, $p = .03$, and for Experiment AF4, $\beta = .27$, $t(57) = 2.03$, $p < .05$. PLO performance was not related to DEMPC situation awareness in Experiments AF1 and AF2, $t(57) = 1.48$ and $t(57) = -.86$, respectively.

The effect of experiment indicates that the mean levels of individual performance changed significantly across experiments. Further analyses revealed that a marginal change in mean values of performance occurred only for PLOs (AF3> AF1 > AF4> AF2). Due to the moderating effect of experiment on the relationship between DEMPC situation awareness and role performance, the role main effects of situation awareness are not interpreted here.

Table 84
Results of the MANOVA Examining the Relationship Between Role Situation Awareness and Role Performance

Source	Num *df*	Den *df*	*F*	Wilks Lambda
Exp	9	107.24	1.78*	.71
AVO SA	3	44	.94	.94
PLO SA	3	44	.61	.96
DEMPC SA	3	44	.01	1.00
Exp*AVO SA	9	107.24	1.31	.78
Exp*PLO SA	9	107.24	1.00	.82
Exp*DEMPC SA	9	107.24	2.32**	.65

*$p<.10$. **$p<.05$

These results indicate that in two of the four experiments the DEMPC's situation awareness has been shown to impact the performance of PLO, such that PLO's performance can suffer as a result of a DEMPC with poor situation awareness.

Summary. Overall, these analyses show that the more individuals with good situation awareness that are on a team, the higher the team performance. However, there is a distinction to be made among roles here. Of the three roles in the UAV-STE the situation awareness of the AVOs is most critical in achieving high levels of team performance. In contrast, the situation awareness of the DEMPCs is important in achieving a good score on the holistic assessment of team situation awareness. High levels of DEMPC situation awareness in some cases also have a positive impact on the performance of the PLO. The relationship between DEMPC situation awareness and team situation awareness has been explained in terms of the DEMPC's specific awareness of the situation awareness query that was repeated regarding number of targets remaining. The role of AVO situation awareness to team performance is interesting and may indicate that it takes situation awareness on the part of the AVO to keep the team moving rapidly from target to target over a 40-minute mission. Without good situation awareness an AVO may not be able to move as quickly or to quickly make changes in accord with others needs.

4.15.3 Individual Taskwork Knowledge

Since taskwork was not measured consistently across all four experiments, several preliminary steps were needed to prepare the data for analysis. First, each experiment had a different number of knowledge sessions. Data from the second knowledge session for each experiment were chosen for the analysis. The rationale for this choice was that by the second knowledge session, participants had had adequate experience with the task, and were not fatigued as they were for knowledge sessions towards the end of each experiment. The exception to this is the fourth experiment that had only a single knowledge session, occurring after the fourth mission, which

was therefore included in the analysis. In addition, data in all four experiments in the following analyses were re-scored with newly devised Pathfinder referents (see measures section of Experiment 1 for more information). In the following analyses only the overall taskwork knowledge accuracy metric was used.

As in previous analyses performance data from Mission 4 were used because it was this mission that represented asymptotic performance levels throughout all four experiments. Univariate analyses of covariance and linear regressions were used to answer the research questions listed above.

Individual taskwork knowledge. These analyses address the first two research questions, namely, how does individual taskwork relate to team performance and to team taskwork measured holistically. A linear regression was used to relate individual taskwork (the maximum and the range of accuracy scores for each team) to team performance and to team taskwork accuracy (a holistic measure of taskwork in which team members reached consensus on taskwork ratings). Data from all four experiments were used in the following analyses. The first regression addressed the relationship between individual taskwork accuracy (maximum and range of taskwork scores on each team) and team performance at Mission 4. The following model was run, where Indmax was the maximum individual accuracy score:

Team Performance = Indmax, Range, Indmax*Range

The analysis revealed that the range and maximum score did not interact to affect team performance, $t(65) = -1.44$. There were no significant main effects of Indmax, $t(68) = 1.31$, or Range, $t(65) = 1.49$, indicating that individual taskwork accuracy was not predictive of team performance.

The second regression examined the relationship between individual taskwork accuracy and team-level taskwork accuracy. The following model was run, where Indmax was the maximum individual accuracy score:

Team Taskwork Accuracy = Indmax, Range, Indmax*Range

The analysis revealed that the interaction between Indmax and Range was not significant, $t(65) = -1.26$. Further, there was no significant main effect of Range, $t(65) = .93$. There was however, a significant main effect of Indmax, $t(68) = 3.40$, $p < .01$, $\beta = .80$, suggesting that teams with an individual who had high individual taskwork accuracy was more accurate in terms of team (i.e., holistic) taskwork accuracy.

Role-specific taskwork knowledge. This section addresses the latter three research questions, that is, how does the taskwork accuracy associated with a particular role (i.e., AVO, PLO, and DEMPC) relate (1) to team performance, (2) to team-level taskwork accuracy, and (3) to role-specific performance?

To determine the impact of each role's taskwork accuracy on team performance, an ANCOVA was performed. The following model was run:

Team Performance = Experiment, AVO Accuracy, PLO Accuracy, DEMPC Accuracy, Experiment*AVO Accuracy, Experiment*PLO Accuracy, Experiment*DEMPC Accuracy

To test for heterogeneity of slopes across the four experiments, we first interpreted the interactions. The Exp*PLO accuracy and Exp*DEMPC accuracy interactions were significant indicating that the relationships between the PLO's and DEMPC's taskwork accuracy and team performance were significantly different across experiments (see Table 85). No significant main effects were found. To further explore the interaction between experiment and PLO accuracy, individual correlations were run for each experiment (see Table 86).

Table 85
Results of the Univariate ANCOVA Examining the Relationship Between Individual Taskwork Accuracy and Team Performance

Source	*df*	*F*
Exp	3, 53	1.07
AVO Acc	1, 53	.09
PLO Acc	1, 53	1.47
DEMPC Acc	1, 53	2.66
Exp*AVO Acc	3, 53	.50
Exp*PLO Acc	3, 53	2.82**
Exp*DEMPC Acc	3, 53	2.29

*$p<.10$. **$p<.05$

Table 86
Results of Correlations of Mission 4 Performance and PLO Accuracy by Experiment

Experiment	Pearson *r*	*p*
AF1	-.48	.14
AF2	-.14	.59
AF3	-.02	.93
AF4	.32	.17

Although there were no significant correlations, Experiments AF1, AF2, and AF3 show that Mission 4 performance is negatively correlated with PLO taskwork accuracy meaning that better performing teams had PLOs that were less accurate. This trend is especially prominent in Experiment AF1. Experiment AF4 showed a positive correlation, as better performing teams had more accurate PLOs.

To further explore the interaction between experiment and DEMPC accuracy, individual correlations were run for each experiment. The results are shown in Table 87. Although there were no significant correlations, Experiments AF1, AF2, and AF4 show that Mission 4 performance was positively correlated with DEMPC taskwork accuracy meaning that better performing teams also had more accurate DEMPCs. This is especially prominent in Experiment AF1. Experiment AF3 showed a negative correlation meaning that better performing teams had DEMPCs that were not as accurate.

Table 87
Results of Correlations of Mission 4 Performance and DEMPC Accuracy by Experiment

Experiment	Pearson *r*	*p*
AF1	.43	.18
AF2	.16	.53
AF3	-.22	.34
AF4	.09	.71

In general, these findings are not surprising given that the taskwork results seem to vary from experiment to experiment, perhaps as a function of the placement of each knowledge session. We have most confidence in the placement of the knowledge session in AF4 (Experiment 2), which suggests a potential positive relationship between the PLO's taskwork knowledge and team performance.

The next research question addresses the effect of each role's taskwork accuracy on team, or holistic, taskwork accuracy. A univariate ANCOVA was used to run the following model:

> Team Taskwork Accuracy = Experiment, AVO Taskwork Accuracy, PLO Taskwork Accuracy, DEMPC Taskwork Accuracy, Experiment*AVO_Taskwork Accuracy, Experiment*PLO_Taskwork Accuracy, Experiment*DEMPC_Taskwork Accuracy

The F-values in Table 88 show that the relationship between PLO taskwork accuracy and team taskwork accuracy were homogeneous ($p \geq .20$) across experiments indicating that PLO accuracy may be predictive of team taskwork accuracy. The correlations for the PLO's taskwork accuracy and team taskwork accuracy were $r = .43, p = .06$ and $r = .51, p = .02$ for Experiments AF3 and AF4 respectively. Correlations for AF1 and AF2 were not significant.

The F-values in Table 88 also show that the relationship between DEMPC taskwork accuracy and team taskwork accuracy were homogeneous across experiments indicating that DEMPC accuracy may also be predictive of team taskwork accuracy. The correlations for the DEMPC's taskwork accuracy and team taskwork accuracy were $r = .66, p < .01$ and $r = .39, p = .09$ for Experiments AF2 and AF4 respectively. Correlations for AF1 and AF3 were not significant. The PLO's and DEMPC's taskwork accuracy bore the strongest relationship to team taskwork accuracy.

Table 88
Results of the Univariate ANCOVA Examining the Relationship Between Role Taskwork Accuracy and Team Taskwork Accuracy

Source	*df*	*F*
Exp	3, 53	.42
AVO TA	1, 53	2.52
PLO TA	1, 53	9.22**
DEMPC TA	1, 53	5.39**
Exp*A_TA	3, 53	.02
Exp*P_TA	3, 53	.39
Exp*D_TA	3, 53	.20

*$p<.05$. **$p<.01$

The impact of each role's taskwork accuracy on performance associated with each role was assessed using a multivariate analysis of variance (MANOVA). The following model was run:

> AVO Performance PLO Performance DEMPC Performance = Experiment, AVO Taskwork Accuracy, PLO Taskwork Accuracy, DEMPC Taskwork Accuracy, Experiment*AVO_Taskwork Accuracy, Experiment*PLO_Taskwork Accuracy, Experiment*DEMPC_Taskwork Accuracy

In testing the heterogeneity of the role taskwork accuracy-role performance relationships, it became clear that there were no relationships between any role's taskwork accuracy and role performance (see Table 89). Role taskwork accuracy does not predict role performance and this relationship did not vary between experiments.

Table 89
Results of the MANOVA Examining the Relationship Between Role Taskwork Accuracy and Role Performance

Source	*df*	*F*
Exp	9, 119	.97
AVO TA	3, 49	.18
PLO TA	3, 49	1.55
DEMPC TA	3, 49	1.89
Exp*A_TA	9, 119	1.07
Exp*P_TA	9, 119	.64
Exp*D_TA	9, 119	.389

*$p<.10$. **$p<.05$

Summary. These analyses in general show little relation between how much an individual (or role) knows about the task and team performance. On the other hand individual knowledge about the task is related to accuracy on the team-level knowledge test. The maximum taskwork score on a team is related to the team-level score in a positive direction. Further, DEMPCs and PLOs tend to have more of an impact on this relationship than AVOs.

4.15.4 Individual Teamwork Knowledge

Since teamwork was not measured consistently across all four experiments, several preliminary steps were needed to prepare the data for analysis. First, each experiment had a different number of knowledge sessions. As mentioned above in the analyses on individual taskwork knowledge, data from the second knowledge session for each experiment was chosen for the analysis. Also Experiments AF1 and AF2 used a different method to measure teamwork knowledge than Experiments AF3 and AF4. Therefore, the following analyses will cover Experiments 3 and 4 only and will use the overall teamwork knowledge accuracy metric. As in the previous analyses, performance data from Mission 4 were used for the performance variable.

Individual teamwork knowledge. These analyses address the first two research questions listed above, namely, how does individual teamwork knowledge relate to team performance and teamwork measured holistically? A linear regression was used to relate individual teamwork

(the maximum and the range of accuracy scores for each team) to either team performance or team-level teamwork knowledge accuracy (a holistic measure of teamwork in which team members reached consensus on the teamwork task).

The first regression addressed the relationship between individual teamwork knowledge accuracy and team performance at Mission 4. The following model was run, where Indmax was the maximum individual accuracy score:

Team Performance = Indmax, Range, Indmax*Range

The analysis revealed that the interaction between Indmax and Range was not significant, $t(36) = 1.07$, indicating that range and maximum score did not interact to affect team performance. There were no significant main effects of Indmax, $t(36) = -1.43$, or Range, $t(39) = -1.05$, indicating that individual teamwork accuracy was not predictive of team performance.

The second regression examined the relationship between individual teamwork accuracy and holistic (team) teamwork accuracy. The following model was run:

Holistic Teamwork Accuracy = Indmax, Range, Indmax*Range

The analysis revealed that the interaction between Indmax and Range was not significant, $t(36) = .86$. Further, there was no significant main effect of Range, $t(39) = -1.03$, or of Indmax, $t(36) = 1.21$, indicating that individual teamwork accuracy was not predictive of holistic teamwork accuracy.

Role-specific teamwork knowledge. This section addresses the latter three research questions, that is, how does the teamwork accuracy associated with a particular role (i.e., AVO, PLO, and DEMPC) relate (1) to team performance, (2) to holistic teamwork accuracy, and (3) to role-specific performance?

To determine the impact of each role's teamwork accuracy on team performance, an ANCOVA was performed. The following model was run:

Team Performance = Experiment, AVO Accuracy, PLO Accuracy, DEMPC Accuracy, Experiment*AVO Accuracy, Experiment*PLO Accuracy, Experiment*DEMPC Accuracy

To test for heterogeneity of slopes across the four experiments, we first interpreted the interactions. The analyses revealed nothing significant (see Table 90) indicating that teamwork accuracy associated with roles was not significantly different across experiments and was not related to team performance.

Table 90
Results of the Univariate ANCOVA Examining the Relationship Between Individual Teamwork Accuracy and Team Performance

Source	*df*	*F*
Exp	1, 32	.02
AVO Acc	1, 32	2.43
PLO Acc	1, 32	.25
DEMPC Acc	1, 32	2.26
Exp*AVO Acc	1, 32	.28
Exp*PLO Acc	1, 32	.01
Exp*DEMPC Acc	1, 32	.53

*$p<.10$. **$p<.05$

The next research question addresses the effect of each role's teamwork accuracy on team, or holistic, teamwork accuracy. A univariate ANCOVA was used to run the following model:

> Holistic Teamwork Accuracy = Experiment, AVO Teamwork Accuracy, PLO Teamwork Accuracy, DEMPC Teamwork Accuracy, Experiment*AVO_Teamwork Accuracy, Experiment*PLO_Teamwork Accuracy, Experiment*DEMPC_Teamwork Accuracy

The F-values in Table 91 show that the relationships between AVO, PLO, and DEMPC teamwork accuracy (main effects) and holistic teamwork accuracy were not heterogeneous across experiments (i.e., interactions between experiment and role taskwork accuracy were not significant). Furthermore, each role's taskwork accuracy significantly predicted holistic teamwork accuracy. The AF3 correlation for the PLO's teamwork accuracy and holistic teamwork accuracy was $r = .49$, $p = .03$. The PLO's correlation for Experiment AF4 was not significant. The correlations for the DEMPC's teamwork accuracy and holistic teamwork accuracy was $r = .54$, $p = .01$ and $r = .60$, $p = .01$ for Experiments AF3 and AF4 respectively. The correlation for the AVO's teamwork accuracy and holistic teamwork accuracy was $r = .56$, $p = .01$ for AF4. The AVO's correlations for Experiment AF3 were not significant.

Table 91
Results of the Univariate ANCOVA Examining the Relationship Between Role Teamwork Accuracy and Holistic Teamwork Accuracy

Source	*df*	*F*
Exp	1, 32	.03
AVO TA	1, 32	6.14*
PLO TA	1, 32	3.48*
DEMPC TA	1, 32	12.88**
Exp*A_TA	1, 32	1.24
Exp*P_TA	1, 32	.64
Exp*D_TA	1, 32	.05

*$p<.01$. **$p<.05$

The impact of each role's teamwork accuracy on performance associated with each role was assessed using a multivariate analysis of variance (MANOVA). The following model was run:

> AVO Performance PLO Performance DEMPC Performance = Experiment, AVO Teamwork Accuracy, PLO Teamwork Accuracy, DEMPC Teamwork Accuracy, Experiment*AVO_Teamwork Accuracy, Experiment*PLO_Teamwork Accuracy, Experiment*DEMPC_Teamwork Accuracy

Across all modeled effects, there were no relationships between any role's teamwork accuracy and role performance (see Table 92). Role teamwork accuracy does not predict role performance and this relationship did not vary between experiments.

Table 92
Results of the MANOVA Examining the Relationship Between Role Teamwork Accuracy and Role Performance

Source	*df*	*F*
Exp	3, 28	.71
AVO TA	3, 28	1.33
PLO TA	3, 28	.91
DEMPC TA	3, 28	1.65
Exp*A_TA	3, 28	1.69
Exp*P_TA	3, 28	.11
Exp*D_TA	3, 28	.83

*$p<.10$. **$p<.05$

Summary. The analyses on the effects of role-specific teamwork knowledge demonstrate that the teamwork knowledge associated with DEMPC's, AVO's, and PLO's can impact team-level teamwork accuracy. Overall however, these analyses show that team performance is not significantly impacted by the teamwork knowledge of individuals or associated with performance of specific roles on the team.

4.15.5 Individual Verbal Working Memory Capacity

Verbal working memory capacity of participants in our experiments was only collected at the beginning of Experiments AF3 and AF4 so this analysis is restricted to these two experiments. We also decided to use Mission 5 performance data because we believed that when the workload increases, working memory capacity, which is the ability to store and manipulate information, should act as an effective predictor of performance. Each member of a team has a score on the working memory task, but there is no corresponding team-level, or holistic measure of working memory capacity. Therefore, our analysis of verbal working memory capacity only addresses Questions 1, 3 ,and 5 concerning the influence of individual and role-specific working memory on team and role-specific performance.

To predict team performance from individual level scores, the following model was run, where Indmax was the maximum value of the individual member's scores on a team and Range was the maximum score minus the minimum score on a team:

Team Performance = Indmax, Range, Indmax*Range

A regression was run using these three variables to predict team performance during Mission 5. None of the variables were significant predictors of team performance, ($t = 1.49$ for maximum score, $t = .94$ for range, $t = -1.08$ for the interaction term).

To address the third question we examined whether the scores associated with specific team roles (i.e., AVO, PLO, or DEMPC) on the verbal working memory task could be used to predict team performance in Mission 5. An ANCOVA was used to run the following model:

Team Performance = Experiment AVO WM, PLO WM, DEMPC WM, Experiment*AVO WM, Experiment*PLO WM, Experiment*DEMPC WM

Note: WM = Working memory

Only the interaction term for AVO working memory and experiment was significant, $F(1, 31) = 2.78, p = .10$), indicating that the relationship between AVO working memory and performance differed in Experiments AF3 and AF4. A test of simple effects revealed no significant effect for AF3, $t = -.60$, but a significant positive effect of AVO working memory on team performance for AF4, $t = 2.15, p < .05, \beta = .45$.

Finally, we also examined whether verbal working memory scores could be used to predict role performance scores on the UAV task. A MANOVA was used to run the following model:

AVO Performance PLO Performance DEMPC Performance = Experiment, AVO WM, PLO WM, DEMPC WM, Experiment* AVO WM, Experiment* PLO WM, Experiment* DEMPC WM

A significant interaction between experiment and DEMPC working memory was obtained that affected PLO performance in Mission 5, $F(1, 31) = 10.40, p < .01$. Tests of simple effects revealed that DEMPC working memory had a significant effect on PLO performance for AF3, $t = 4.31, p < .01, \beta = .72$, but not for AF4, $t = -.56$. Significant main effects were found for AVO working memory and DEMPC performance, $F(1, 31) = 3.50, p < .10$ as well as for DEMPC working memory and for DEMPC Mission 5 performance, $F(1, 31) = 3.22, p < .01$. A regression revealed that both AVO working memory, $t = 2.59, p < .05, \beta = .37$, and DEMPC working memory, $t = 2.43, p < .05, \beta = .35$, were positively associated with DEMPC performance.

To summarize, AVO working memory seems to have an impact on team performance. It also appears that verbal working memory scores are predictive of DEMPC Mission 5 performance. DEMPCs who obtained higher verbal working memory scores performed better in Mission 5, the first high workload mission. Teams whose DEMPCs obtained higher scores during Mission 5 also had AVOs with higher verbal working memory scores. Further, high verbal working memory on the part of the DEMPC also seems to be related to PLO performance.

We decided to examine whether these effects were moderated by dispersion condition (co-located or distributed status). An ANCOVA was used to run the following model:

DEMPC Performance = Dispersion, AVO WM, DEMPC WM, Dispersion*AVO WM, Dispersion*DEMPC WM

No interaction between dispersion condition and AVO working memory was found, $F(1, 33) < 1$, but there was a significant interaction between dispersion condition and DEMPC working memory, $F(1, 31) = 3.30, p < .10$. Follow-up tests revealed a significant effect of DEMPC working memory score on DEMPC Mission 5 performance for co-located teams, $t = 2.79, p < .05, \beta = .55$, but not for distributed teams, $t = .28$.

Summary. These results are interesting and support a finding reported in the appendix on workload measures that co-located DEMPC perceive greater workload demands than distributed DEMPCs. Specifically, the relation between DEMPC working memory and DEMPC performance seems to be present for the co-located teams who perceive high levels of workload demand. Overall, these analyses suggest that verbal working memory capacity plays a role in individual and (to a lesser extent) team performance.

4.15.6 Individual Processing Speed

We decided to use Mission 4, the last low workload mission when teams reached asymptotic levels of performance, to examine whether scores on the processing speed task were predictive of team or role performance on the UAV task. Processing speed measures assess how quickly participants can execute an over learned response, such as deciding whether two simple words have the same meaning, and may reflect one's ability to engage in the quick, effortless type of processing that is characteristic of skilled performance. We hypothesize that processing speed should become a more important predictor as skill increases. Data on the processing speed measure were only collected for Experiment AF4, so experiment was not used as a predictor in our analyses. Each member of a team had a score on the processing speed measure, but there was no corresponding team-level, or holistic measure, so this analysis uses speed to predict team and role performance only. Thus we address Questions 1, 3, and 5 in this section.

To predict team performance, the following regression model was run, where Indmax was the maximum individual processing speed score:

Team Performance = Indmax, Range, Indmax*Range

None of the variables were significant predictors of team performance, $t = -.35$ for maximum score, $t = -.15$ for range, and $t = .28$ for the interaction term.

We also examined whether the scores associated with team member role on the processing speed task could be used to predict team performance in Mission 4. Using ANCOVA, the following model was run:

Team Performance = AVO Process Speed, PLO Process Speed, DEMPC Process Speed

None of the speed measures were significant predictors, $F(1, 16) < 1$, for AVO processing speed, $F(1, 16) < 1$, for DEMPC processing speed, and, $F(1, 16) < 1$, for PLO processing speed.

Finally, we examined whether processing speed scores could be used to predict role scores on the UAV task. A MANCOVA was used to run the following model:

> AVO Performance, PLO Performance, DEMPC Performance = AVO Process Speed, PLO Process Speed, DEMPC Process Speed

No significant results were obtained.

Summary. To summarize, we found no significant effect of processing speed on either team performance or role performance during Mission 4.

4.15.7 Individual Voice Stress

Data presented in this section should be considered preliminary because it is based on only a few cases. The purpose of this analysis is to explore the potential of using voice frequency as a means of assessing workload or stress on-line (i.e., during mission performance). This application of voice analysis is promising, to the extent that voice frequency relates to team performance.

Vocal frequencies, which may measure stress levels, may increase when task difficulty increases. We collected frequency data on male and female team members in Experiments AF3 and AF4 for the first five minutes of the first mission (Mission 1). Five teams from AF3 and two teams from AF4 were available for analysis. We used the first mission because workload should be high when teams first encounter the task. We used median frequencies because the data were skewed. Each member of a team had a frequency value, but there was no corresponding team-level, or holistic measure of frequency. Therefore, our analysis uses the frequency measure to predict team or role performance only (Questions 1, 3, and 5). Experiment was not used as a variable because there were only 2 cases (teams) for AF4.

As stated earlier, software manufactured by Avaaz Innovations was used to obtain the frequency measurements. The sampling frequency was 48 kHz. Several parameters that were recommended by the software manufacturer for voiced speech were used, including 5 db for the silence threshold, 1500 Hz for the Zero-Crossing Frequency Threshold, and 400 Hz for the maximum frequency detected. An approach that initially estimates frequency by making a pass through the first few glottal cycles was used. The default waveform matching method was chosen as the pitch extraction algorithm.

Because frequencies for males were lower (M = 154.76, SD = 55.02, N=16) than for females (M =246.32, SD = 24.74, N =5), we used z-scores that provided a common scale for all team members and allowed us to calculate a team score on each variable. Frequency scores that were standardized separately for males and females across both experiments were used in all of the analyses.

The following regression model was run where the maximum frequency and range of frequencies on the team were used to predict team performance:

Team Performance = Indmax, Range, Indmax*Range

None of the variables were significant predictors of team performance, $t = .75$ for maximum score, $t = 1.27$ for range, $t = -1.30$ for the interaction term. We also examined whether the frequency values associated with specific team roles could be used to predict team performance in Mission 1.

Team Performance = AVO Voice Freq, PLO Voice Freq, DEMPC Voice Freq

The ANCOVA revealed no significant effect of AVO frequency, $F(1, 3) = .47$, DEMPC frequency, $F(1, 3) < 1$, or PLO frequency, $F(1, 3) < 1$, on team performance.

Finally, we examined whether frequency scores could be used to predict role performance scores. A MANCOVA was used to run the following model:

AVO Performance PLO Performance DEMPC Performance = AVO Voice Freq, PLO Voice Freq, DEMPC Voice Freq

The only effect that approached significance was for DEMPC frequency during Mission 1 and DEMPC performance for Mission 1, $F(1, 3) = 5.41$, $p = .10$. Follow-up analysis revealed that DEMPCs with higher frequencies obtained lower scores on the first mission, $t = -1.8$, $p < .13$, $B = -.63$. The correlation between DEMPC frequency at the beginning of Mission 1 and DEMPC performance during Mission 1 was -.63.

Summary. Unfortunately, in our sample, role and gender were confounded. All of the DEMPCs were male whereas two of the AVOs and three of the PLOs were female. When scores on the frequency measure were correlated with performance, males produced a correlation of -.40 whereas females obtained a correlation of .70. Therefore, the negative correlation for DEMPCs may have been due to the gender composition of the sample rather than the difficulty of the role. In general, it would probably be more informative to look at deviations in individual voice frequency, rather than absolute differences across individuals. Nonetheless, we consider that with much larger samples of voice and individuals that these results provide some indication that voice stress may serve as an on-line measure of workload or stress.

4.15.8 Individual Subjective Workload

In our studies NASA TLX ratings of subjective workload were collected in two experiments, Experiment AF3 and Experiment AF4. Thus, TLX data and corresponding performance scores from only these two experiments are used in this archival analysis.

The regression analysis was performed in order to address the first main research question, namely, how individual subjective perceptions of workload (measured by the NASA TLX) relate

to team performance. Subjective workload was not measured at the team level so in this analysis we do not consider team subjective workload (i.e., we address Questions 1, 3, and 5). The differences between Mission 5 and Mission 4 for TLX estimates and corresponding performance scores were entered into this analysis. This metric reflects sensitivity to the workload manipulation, which occurred between Mission 4 (low workload) and Mission 5 (high workload). Thus in this analysis we focus not on absolute impressions of workload which may vary radically from individual to individual, but with the relative deviation of judgments between low and high workload missions. Some individuals may perceive a greater increase in workload between the low and high conditions than others and it is of interest as to whether this perceived change is predictive of performance changes between high and low workload missions.

A regression analysis was performed with the maximum individual TLX difference and the range of total TLX differences obtained for each team as predictors and team performance as the criterion.

The results shown in Table 93 indicate that there are no significant relationships between workload changes reflected by individual TLX and changes in team performance.

Table 93
Results from the Regression Analysis

Source	*SS*	*df*	*MS*	*F*
Max	550	1	550	.12
Range	478	1	478	.11
Max*Range	729	1	729	.16
Total	168,796	39		

To determine the impact of each role's TLX on team performance, a univariate analysis of covariance (ANCOVA) was run. The following model was used:

> Team Performance = Experiment, AVO_TLX, PLO_TLX, DEMPC_TLX, Experiment* AVO_TLX, Experiment* PLO_TLX, Experiment* DEMPC_TLX

First, the heterogeneity of slopes across two experiments was tested to identify differences in role-team relationship across experiments. Looking at the results in Table 94 we can see that there are no differences in the relationship between role TLX and team performance across Experiments AF3 and AF4.

Since the relationship between role TLX and team performance are consistent across experiments, the next step is to examine more specifically this relationship. The same univariate ANCOVA answers this question. No significant relationship appeared between role TLX and team performance (see Table 94). The presented results indicate that sensitivity to change in workload as measured by TLX associated with particular roles is not predictive of change in team performance.

Table 94

Results of the Univariate ANCOVA Examining the Relationship Between Role TLX and Team Performance.

Source	*SS*	*df*	*MS*	*F*
Exp	27	1	27	.01
AVO_TLX	10	1	10	.00
PLO_TLX	6,176	1	6,176	1.35
DEMPC_TLX	2,214	1	2,214	.48
Exp*AVO	1,802	1	1,802	.37
Exp*PLO	6,516	1	6,516	.55
Exp*DEM	195	1	195	.04

To explore the relationship between role-specific TLX sensitivity and role performance a multivariate analysis of variance was performed. The following model was used for this analysis:

> AVO Performance PLO Performance DEMPC Performance = Experiment, AVO_TLX, PLO_TLX, DEMPC_TLX, Experiment*AVO_TLX, Experiment* PLO_TLX, Experiment* DEMPC_TLX

First, the heterogeneity of relationships between role TLX and role performance across experiments was tested using MANOVA. These tests did not reveal any significant differences in these relationships. Also the MANOVA revealed that there are no significant relationships existing between role TLX and role performance (see Table 95).

Table 95
Results of the MANOVA Examining the Relationship Between Role TLX and Role Performance across Experiments.

Source	Num *df*	Den *df*	*F*	Wilks' Lambda
AVO_TLX	3	30	.93	.91
PLO_TLX	3	30	.09	.99
DEMPC_TLX	3	30	.84	.92
Exp*AVO TLX	3	30	.32	.97
Exp*PLO TLX	3	30	1.01	.91
Exp*DEMPC TLX	3	30	.64	.94

Summary. The archival analyses revealed that deviations from low to high workload missions in individual TLX estimates are not related to deviations in team performance for those missions. Also, no significant effects were found associated with a particular role's TLX estimates and a team or role's performance. Subjective workload estimates may not be sensitive to actual workload deviations, which in turn affect performance.

4.15.9 Individual Grade Point Average

Grade Point Average (GPA) was requested of all participants for all four experiments. Four participants did not provide this information, and were excluded from the analyses. There was no team-level GPA and so Questions 1 and 3 are the focus of this analysis. Again, Mission 4

performance was used as an estimate of individual and team performance across the experiments because Mission 4 was the point at which individuals and teams reached asymptotic levels of performance.

To address the question about individual GPA and team performance, a regression was conducted using the maximum and range of GPA among the three team members as the independent variables predicting team performance. Non-significant results were obtained for maximum GPA, $t(62) = .61$ and range of GPA's, $t(62) = .55$. The interaction effect was not tested.

The next analysis addresses the third research question. Specifically, how does the GPA associated with a particular role (i.e., AVO, PLO, and DEMPC) relate to team performance?

To determine the impact of each role's GPA on team performance, a univariate analyses of covariance (ANCOVA) was performed. The following model was run:

> Team Performance = Experiment, AVO_GPA, PLO_GPA, DEMPC_GPA, Experiment*AVO_GPA, Experiment*PLO_GPA, Experiment*DEMPC_GPA

To test for heterogeneity of slopes across the experiments, we first interpreted the interactions. The F-values in Table 96 show that one interaction effect was significant, suggesting that the relationship between the AVOs' GPA and team performance was significantly different across experiments. To determine the differences between experiments, correlations were performed between AVO GPA and team performance for each individual experiment. Only in Experiment AF4 did AVO GPA significantly predict team performance, $r(57) = .65$, $p < .01$. Also, as indicated in Table 96 there was not a significant experiment effect for team performance across experiments. Finally, of the three roles, significant main effects were observed for AVOs and PLOs. However, we interpret the main effect of AVO GPA with caution due to the interaction effect based on AF4 described above.

Table 96
Results of the Univariate ANCOVA Examining the Relationship Between Role GPA and Team Performance

Source	*df*	*F*
Exp	3	1.87
AVO GPA	1	4.08**
PLO GPA	1	16.80***
DEMPC GPA	1	1.06
Exp*AVO GPA	3	2.62*
Exp*PLO GPA	3	.46
Exp*DEMPC GPA	3	1.62

***$p<.01$ ** $p\leq.05$ *$p\leq.10$

Summary. In sum, although individual GPA of the members of a team has no general relationship to team performance in the UAV-STE, the GPA associated with the AVO and PLO roles are related to team performance.

4.15.10 Demographics and Team Composition

Participants in all four experiments provided demographic data regarding gender, major course of study, class standing, level of aviation training, and rank in the military (if applicable). In addition, participants in Experiments AF2 through AF4 indicated their ethnicity. Three participants did not indicate gender, aviation training, or major. Four did not indicate class. Table 97 contains the demographic characteristics for the 207 participants in the four experiments. These data were examined in terms of individual characteristics and team composition.

Table 97
Demographic Characteristics of Participants in Experiments AF1 through AF4

Exp	Gender		Military		Aviation Training		Major		Ethnicity			Class	
	M	F	Yes	No	Yes	No	Non-Tech	Tech	Cauc	Hisp	Othr	Und	Uppr
AF 1	22	8	29	4	5	25	15	15	NA	NA	NA	11	18
AF2	40	14	54	0	4	50	11	43	26	25	3	27	27
AF3	39	21	20	40	5	55	27	33	33	16	11	26	34
AF4	60	0	1	59	3	57	26	34	26	20	14	40	20
Total	161	43	104	103	17	187	79	125	85	61	28	104	99

NA= Not available; data not requested for Experiment AF1.

For the archival analyses of these variables, some initial steps were required to prepare the data. While two variables (i.e., gender, aviation training) were dichotomous, a variety of responses were present for the remaining variables. Participants indicated the rank they held in the military. A new variable was created to indicate if the participant was in the military or not. For ethnicity, a majority of the participants were Caucasian or Hispanic, with the remaining indicating a variety of other ethnicities. The new variable was coded for Caucasian, Hispanic, or Other. Five values were possible for class standing. The derived variable indicated whether the participant was an under-classman (i.e., freshman, sophomore) or an upper-classman (i.e., junior, senior, graduate student). Many responses were supplied for the major course of study. For the analyses, the major was coded as technology-oriented (i.e., mathematics, science, engineering) and other than technology-oriented (e.g., liberal arts, business, education). For the primary analyses, Mission 4 performance, the point at which the teams reached asymptotic levels, was used as an estimate of team performance across the experiments.

Demographic data and team performance. These analyses address the question of the relationship between the demographic variables and team performance. The second research question above was not addressed because there was not an associated team-level characteristic.

Similar to the archival analysis of situation awareness, a numerical variable for each demographic characteristic was created, and was used to perform a correlational analysis with team performance. These variables included counts of males on the team, team members with aviation training, team members in the military, team members of the same ethnicity, upper-classmen on the team, and team members in technical fields. One significant correlation emerged. As the number of team members in the military increased, team performance decreased, $r\,(67) = -.20, p < .10$, n = 69.

Role-specific effects. To determine the impact of each role's demographic characteristics on team performance, Chi Square analyses were conducted for all demographic variables (using a median split on each variable) by role. One significant result emerged. Teams in which the AVO was a male exhibited significantly better team performance than teams with a female AVO, $\chi^2\,(1) = 5.23, p < .10$.

Chi Square analyses were also conducted to assess the impact of each role's demographic characteristics on individual performance. Two significant results were found. AVOs with aviation training performed significantly better than AVOs without aviation training, $\chi^2\,(1) = 3.08, p < .10$. Because there were only three AVOs with prior aviation training, all of whom had performance scores greater than the median, this result must be interpreted with caution. Also, DEMPCs majoring in fields classified as other than technical performed significantly better than those with technical majors, $\chi^2\,(1) = 3.10, p < .10$.

Team composition and team performance. In addition, to assess the effects of team composition, dichotomous variables were created to indicate the composition of the team. For gender, mixed and same gender teams were identified. Teams consisting of members with similar majors (e.g. all liberal arts majors, all engineering majors) were differentiated from teams with at least one member with a major in a different field. For the military rank variable, the two groups consisted of teams dominated by members in the military and teams that were not. Teams with two or three members in the military were considered to be dominant military. Teams with at least one member with aviation training were differentiated from teams in which no one had received such training. For class standing, the sample was divided between teams that had at least one senior or graduate student and those that did not. Finally, the teams were divided based on ethnic composition, and consisted of groups in which all members were of the same ethnicity, and those that had at least one member of a different ethnicity. Table 98 contains the demographic composition of the 69 teams that participated in the experiments.

For all of the analyses, the median score was calculated for the team performance scores for the mission, and was used as the cut-off for low and high scoring teams. The number of low and high scoring teams for each value of the derived dichotomous variables was used in Chi Square analyses. Across experiments, no significant relations emerged between the team composition variables in Table 98 and team performance.

Table 98
Demographic Composition of Teams in Experiments AF1 Through AF4

	Gender		Military		Aviation Training		Major		Ethnicity		Class	
	Mix	Same	0 - 1	2 - 3	None	1+	2+ Sim	None Sim	Same	Mix	All F/So/J	Sr/ GS
AF1	5	5	0	11	7	4	5	5	NA	NA	9	2
AF 2	9	9	0	18	15	3	13	5	7	11	7	11
AF3	14	6	17	3	15	5	10	10	4	16	11	9
AF4	0	20	20	0	17	3	12	8	0	20	7	13
Total	28	40	37	32	54	15	40	28	11	47	34	35

NA= Not available; data not requested for Experiment 1.

Team perceptions. There was one other measure taken at an individual level that can also be explored. At the conclusion of Experiment AF3, participants rated to what extent they agreed or disagreed with statements about the experiment. Specifically, the statements addressed the participants' enjoyment and performance during the study, as well as their perceptions of their teammates' tasks and performance. All items were rated on a scale of 0 (disagree) to 4 (agree).

Chi Square analyses were conducted to assess the relationship between gender composition and team performance across missions. An initial exploration of the data hinted at potential differences between mixed and same gender teams for Experiment AF3. These differences did not surface in Experiments AF1 and AF2. Specifically, in Experiment AF1, Chi Square analyses indicated that the number of mixed and same gender teams in the groups of high performing and low performing teams (as determined by a median split) did not differ from what we would expect by chance. This finding was consistent at each mission and when considering performance averaged across all missions. However, although the observed Chi Squares were not significant, it was typically the case that more same gender teams were in the high performing group than in the low performing group. For Experiment AF2, the number of same and mixed gender teams in the high and low performing groups also did not deviate significantly from what we would expect by chance. In contrast, in Experiment AF3, there were significantly more same gender teams in the high performing group than what we would expect by chance in Missions 1 and 2, χ^2 (1) = 3.81, p<.10, for both missions. Significant differences were not found for the remaining missions. These analyses were not conducted for AF4, as all teams were composed of males.

Composite scores were computed for participant enjoyment, performance perception, and participant perception of the team member in each of the three roles by summing the ratings for all items in each composite. The items included in each composite score are listed in Table 99. Chi Square analyses were performed by splitting each composite measure at the median and juxtaposing high and low composite scores against other variables including co-located and

distributed teams, military and non-military team members, and males and females. The remaining demographic variables were not included in these analyses due to the uneven distribution of participants. There were no significant interactions between rating score and dispersion or between rating score and gender. One significant result, however, emerged from the analyses of military status. Military affiliated team members were significantly more likely than their non-military counterparts to rate their performance higher χ^2 (1)= 3.42, p<.10. This is interesting given the negative correlation between number of team members with a military background and team performance.

Table 99
Rated Items Used to Derive Non-demographic Debriefing Measures.

Composite Measures	Items
Participant Enjoyment	I enjoyed participating in this study
	I enjoyed the team task part of this study
	I would welcome the opportunity to participate in this study in the future
	I would like to work with my fellow team members again
	I like playing video and computer games
	I like to be part of a team
Performance Perception	I was a successful member of the team
	My team worked well together
	I performed well on this task
	My team performed well on this task
	My individual performance is important to our team
	Performance was evaluated at the individual level
	Performance was evaluated at the team level
Participant perception (AVO)	The AVO was competent
	The AVO contributed to the team
	The AVO tried hard
	The AVO was lucky
	The AVO had an easy task
	The AVO was likable
Participant perception (PLO)	The AVO was competent
	The PLO contributed to the team
	The PLO tried hard
	The PLO was lucky
	The PLO had an easy task
	The PLO was likable
Participant perception (DEMPC)	The DEMPC was competent
	The DEMPC contributed to the team
	The DEMPC tried hard
	The DEMPC was lucky
	The DEMPC had an easy task
	The PLO was likable

Summary. Overall, these analyses indicate that individual demographics and team composition, regardless of the variable, had little or no effect on team or individual performance. Several relatively weak differences, however, were observed. First, teams with more members in the military exhibited poorer team performance than teams consisting of fewer military members. Interestingly they were also more likely to rate their performance as better than their nonmilitary counterparts. Second, several differences were found concerning participant roles. The results suggested that prior aviation training may have had a positive effect on an AVO's individual task performance. Another significant finding concerned the gender of the AVO. While males did not perform significantly better in terms of individual performance scores, team performance was significantly better when the AVO was a male. In addition, analyses of individual DEMPC performance indicated that those majoring in fields that were not technically-oriented performed better than their technically-oriented counterparts. Finally, initial team performance for same gender teams was significantly better than for mixed gender teams in the first two missions of Experiment AF3, but by Mission 3, no differences were evident.

4.16 Archival Analysis of Individual and Role-Associated Factors: Discussion

The purpose of the archival analyses presented in this section was to *investigate the relation between individual characteristics and team cognition and performance.* Using the largest sample available of up to 69 three-person teams several characteristics of individuals were investigated. These included individual performance on the UAV-STE task, individual knowledge-related variables (situation awareness, taskwork knowledge, teamwork knowledge), individual cognitive processing variables (verbal working memory capacity, processing speed), physical variables including voice frequency, demographic variables, and a variety of judgment variables (NASA TLX, task and performance ratings).

Individual and team performance results confirmed that the UAV-STE is an interdependent task in which all three team members are crucial to team performance. High scoring teams tended to be composed of high scoring individuals with performance of no single role driving team performance.

Of the cognitive variables tested, taskwork and teamwork knowledge of individuals was not related to team or role-specific performance. However, these factors were important to the holistic taskwork and teamwork judgments. Teams who had knowledgeable team members obtained higher scores on the consensus-based tests. The lack of a strong relation between individual knowledge of taskwork or teamwork and team performance is not unexpected, given the relatively weak and sporadic correlations between team knowledge and team performance. Although it is clear that knowledge of taskwork and teamwork are important at some level, it appears that the level that must be achieved to meet the training criterion is sufficient and that improvements in knowledge beyond this point do not necessarily map onto improved performance. Instead it seems that the best teams focus more on team coordination or process, rather than improved knowledge.

Alternatively, situation awareness as measured by the repeated query did seem relevant to team performance. The more individuals on a team with good situation awareness, the higher the team situation awareness and team performance. Further there seem to be particular roles for

which this factor is more critical. The situation awareness of the AVO is related to team performance and the situation awareness of the DEMPC is related to team situation awareness and PLO performance. So, in spite of the fact that we are not sure that our situation awareness measure is measuring what we mean by team situation awareness, whatever it is measuring is related to team performance. It is perhaps picking up on a very specific type of situation awareness that has to do with awareness of the experimental situation. The situational information that might be acquired with experience is the number of targets that can be expected for each mission and the criterion for success within the experiment and situation awareness test. People who are good at this, especially AVOs and DEMPCs, tend to be on high-performing teams.

Of all of the other variables tested, the ones that were most relevant to team performance were working memory capacity and grade point average. Some demographic and team composition factors also seemed somewhat related, but the results are too sporadic to draw any firm conclusions. In regard to working memory capacity and grade point average, the relationships between the characteristic and team performance is role-specific. Working memory is important for DEMPCs and AVOs, not PLOs, whereas grade point average is important for AVOs and PLOs, not DEMPCs. These role-specific patterns make sense in relation to the task. For instance, the navigation and planning performed by DEMPC and AVO require some memory for where the team has been and where the team is going. In the sense that these variables are measures of innate cognitive ability, certain people may be better suited to certain roles.

Finally, the significance of this analysis resides in the fact that it is applied to heterogeneous teams in which roles are different, though interdependent. This raises a host of new questions when it comes to individual differences. It is not simply the case that a characteristic of an individual is important (or not) for effective team performance, but as shown here, the presence or strength of this relation may depend on the team role. This information has implications for team composition and training. If working memory is critical for some roles, but not for others, then individuals could be assigned to team roles on the basis of their working memory capacity. Low working memory individuals could be assigned to be the PLO, rather than the DEMPC or AVO. Also, individuals assigned to specific roles may be trained differently. For instance, based on our findings in the UAV-STE, it may be a good idea to give the DEMPC and AVO specific training or decision aids to facilitate situation awareness. Overall these results provide further evidence that for heterogeneous teams like these the whole is not simply the sum of the parts.

In the next and final section of this report we provide a similar archival analysis aimed at evaluating the validity and reliability of our measures.

4.17 Archival Analysis to Evaluate Measures

The purpose of this section is to address the fourth and final objective of this project which is to evaluate the newly developed measures and metrics of team cognition in terms of reliability and validity through an archival analysis on data from four previously conducted CERTT UAV-STE experiments. Thus the approach for this part of the project is similar to that of the last section. Here we conduct archival data analyses of four CERTT UAV-STE experiments but instead of

identifying individual and role-related characteristics relevant to team performance, we evaluate our measures.

The specific tasks involved in this part of the effort are: (1) Assemble data collected from four CERTT-UAV studies, (2) Evaluate across the four studies measures of team cognition, especially in terms of measure reliability and validity, (3) Conduct multi-trait multi-method (MTMM) analysis on data collected from Experiment 1, (4) Examine the benefit of holistic vs. collective measures of team cognition across the four studies, and (5) Address aggregation of individual data for measures of team cognition at the collective level.

In the following section we examine our primary measures which include measures of team performance, team process, situation awareness, taskwork knowledge, and teamwork knowledge. We evaluate each measure (or family of measures) in terms of reliability and validity. Reliability is addressed through a test-retest paradigm and in terms of consistency of findings across studies. Validity is addressed using a regression analysis with the various measures as predictors and team performance as a criterion. We also apply the MTMM approach to the examination of validity.

Also in this section we examine in greater detail our holistic knowledge measures, as these have been one of the central measurement innovations of this effort. We evaluate whether collective vs. holistic measurement makes a difference and whether there is differential validity of one type of measure over the other. We originally speculated that collective measures of team knowledge did not capture team interaction or process and therefore contend that holistic measurement would be more appropriate for heterogeneous teams where process is more complex than simple aggregation schemes. Finally, we look closely at our holistic measures in terms of the aggregation schemes used by our teams to reach consensus. We then examine whether there is information in this process that is relevant to team performance.

4.18 Archival Analysis to Evaluate Measures: Methods

In this section we use data collected from the same four studies referred to in the previous section. We again use the convention of referring to the studies as AF1, AF2, AF3, and AF4 where AF1 and AF2 are experiments conducted in a previous effort and AF3 and AF4 correspond to Experiments 1 and 2 of this effort. Participants and experimental methods are described in the previous section. Measures are described in the Experiment 1 section on primary measures.

4.19 Archival Analysis to Evaluate Measures: Results

4.19.1 Reliability

We used two methods to assess reliability of our measures. First, we examined the pattern of findings between the different studies. In this approach, replicated findings for a measure across studies indicate reliability. To the extent that the measure fails to replicate, reliability is weakened.

Second, we examined test-retest reliability by using a multiple degree of freedom repeated measures test to find differences among equivalent missions. This second approach is very similar to computing an intra-class correlation coefficient. A detectable difference among the equivalent missions indicates unreliability of the measure. In order to justify attempting to accept a null hypothesis, we set alpha for these tests at a high value of .2. A *p*-value of less than .2 indicates that at least one of the putatively equivalent missions is detectably different from the others, and so the measure's reliability is weakened. Equivalent missions (or knowledge sessions) were those missions (or knowledge sessions) after which the team has reached asymptote and those within the same difficulty condition. We used the performance acquisition curve to define an approximate performance asymptote between Missions 4 and 5.

The lack of a true asymptote makes it difficult to implement this measure of reliability. On the other hand, it is difficult to justify continuing a costly research study for one additional mission, knowing that the missions are simply to be considered equivalent. Equivalent missions for AF1 include Missions 5, 6, 7, 9 and 10. Mission 8 is excluded because teams were re-learning after their break. AF2 and AF4 each included five missions, and because teams could legitimately improve between Missions 4 and 5, these missions do not constitute an adequate test of reliability. For AF3, teams may legitimately continue learning between Missions 5 and 6, which are the first of three high workload missions. Also, there was a communication "glitch" manipulation introduced only at Mission 6. Therefore, we will test performance, process, and situation awareness reliability by comparing Missions 5, 6, 7, 9, and 10 of AF1.

Taskwork and teamwork knowledge measures present a different case because they are measured in sessions apart from missions. There was only one session in AF4 and in AF3 the sessions were placed before all missions and after all missions so change between sessions was expected. AF1 had four knowledge sessions and AF2 had three. The first session for AF1 was after the first mission and the first session for AF3 was immediately after training. In both cases we would expect to see change between Session 1 and Session 2. The last sessions of AF1 and AF2 may also be different due to the fact that they occurred at the end of a long study and participants may have been fatigued or simply tired of the same repeated task. This leaves AF1 Sessions 2 and 3 as the test case for teamwork and taskwork. However, the teamwork measure has evolved considerably over the course of the four experiments so the measure that was used in AF1 is not the same as the measure used in AF3 and AF4 in this effort. Therefore we will not be able to examine the test-retest reliability of teamwork knowledge using this method.

Performance. First, in Table 100 we report descriptive statistics for our measure of team performance across the four experiments which are analyzed in this section, as well as the benchmarking study (AF5).

There is a problem with performance reliability assessment. The definition of asymptote is based on sequential pairwise tests of performance. This, of course, renders the test of performance reliability primarily circular. Nevertheless, we present the findings to show the stability of performance after asymptote.

Table 100
Descriptive Statistics of Team Performance for each Experiment

Experiment	N	Min	Max	*M*	*SD*	Var
AF1	106	-.93	639.49	403.00	130.74	17092.44
AF2	90	30.07	584.37	380.33	118.17	13963.06
AF3	140	63.21	539.11	358.24	76.43	5842.04
AF4	100	174.12	580.71	370.31	94.74	8976.23
AF5	25	223.54	616.64	430.95	126.69	16051.42

N is based on teams x missions.

In examining test-retest reliability of AF1 Missions 5, 6, 7, 9, and 10, our team performance measure showed mission-wise changes under $\alpha = .20$, $F(4, 37) = 3.12$, $p = .03$. There was improvement between Mission 7 ($M = 431.81$, $SE = 16.64$) and Mission 9 ($M = 504.13$, $SE = 17.71$), $F(1, 37) = 1.98$, $p = .17$. This may be explained by the fact that Missions 7 and 10 were not the same task scenario as the other missions (including Mission 9), though differences were designed to be merely superficial. That is, critical factors such as number of targets and ROZ boxes were kept constant.

There is also positive support for the reliability of our performance measure based on replicability of results over experiments. First, performance seems to asymptote between Mission 4 and 5 for all testable studies. Second, for AF2, AF3 and AF4, means show a replicated, but not statistically detectable, advantage for distributed/no sharing teams over co-located/sharing teams (see Table 101).

Table 101
Consistent but not Detectable Performance Advantage for Distributed/Non-shared Teams.

	Condition	*M*	*SD*	N
	Non-shared	388.79	125.14	40
AF2	Shared	373.56	113.11	50
	Total	380.33	118.17	90
	Co-located	357.62	80.49	70
AF3	Distributed	358.86	72.73	70
	Total	358.24	76.43	140
	Co-located	366.19	95.58	50
AF4	Distributed	374.43	94.68	50
	Total	370.31	94.74	100

Further, referring to Table 102, the last low workload mission (Mission 4) is consistently better at performance than the first high workload mission (Mission 5), for both AF3, $F(1, 19) = 24.87$, $p < .01$, and AF4, $F(1, 19) = 29.60$, $p < .01$. We conclude that the performance measure has adequate reliability.

Table 102
Means and Standard Deviations for Last Low Workload Mission and First High Workload Mission, for AF3 and AF4

Mission	AF3		AF4	
	M	*SD*	*M*	*SD*
4	432.65	62.66	427.87	95.74
5	358.19	49.88	347.14	72.03

Team process. In Tables 103 and 104 we report descriptive statistics for our two measures of team process (critical incident process and summary process) across the four experiments we analyze in this section, as well as the benchmarking study (AF5). Summary process was only collected under the current effort.

Table 103
Descriptive Statistics of Critical Incident Process for each Experiment

Experiment	N	Min	Max	*M*	*SD*	Var
AF1	107	0.06	1.00	0.77	0.17	0.03
AF2	90	0.17	1.00	0.69	0.20	0.04
AF3	140	0.20	0.90	0.56	0.16	0.02
AF4	100	0.10	1.00	0.52	0.18	0.03
AF5	25	0.44	0.90	0.72	0.13	0.02

N is based on teams x missions.

Table 104
Descriptive Statistics of Summary Process Ratings for each Experiment

Experiment	N	Min	Max	*M*	*SD*	Var
AF3	140	1.50	4.88	3.28	0.74	0.55
AF4	100	1.25	5.00	3.36	0.96	0.93
AF5	25	2.88	5.00	4.29	0.67	0.45

N is based on teams x missions.

The reliability tests of the process measures also rest on a comparison of Missions 5, 6, 7, 9, and 10 of AF1. Critical incident process, the only process measure used in AF1, was adequately reliable $F(4, 38) = 1.51, p = .22$.

Next we consider reliability in terms of replicating findings across studies. In AF4, both summary process and critical incident process were positively related to performance. Summary process was positively predictive of performance for distributed teams at Mission 5 in AF3, $F(1, 8) = 3.72, p = .09, \beta = .56$. This positive relationship was replicated in AF4, $F(1, 8) = 16.95, p < .01, \beta = .82$. Also, critical incident process was positively predictive of performance for distributed teams at Mission 4 in AF3, $F(1, 8) = 27.60, p < .01, \beta = .88$. Also, for both AF3 and AF4, and to a lesser extent for AF1, process shows improvement in early missions, and (for AF3 and AF4) a decline with increases in workload. Therefore we conclude that both of our process measures have adequate reliability.

Situation awareness. In Table 105 we report descriptive statistics for the overall accuracy metric for the repeated situation awareness query across the four experiments, which we analyze in this section as well as for the benchmarking study (AF5).

Table 105
Descriptive Statistics of Situation Awareness Accuracy to the Repeated Query for each Experiment

Experiment	N	Min	Max	*M*	*SD*	Var
AF1	81	0	3	1.51	1.32	1.75
AF2	81	0	3	.90	1.19	1.42
AF3	140	0	3	.78	.97	.94
AF4	100	0	3	1.04	1.17	1.37
AF5	24	2	3	1.37	1.28	1.64

N is based on teams x missions.

The test-retest analysis of the situation awareness measure also rests on a comparison of Missions 5, 6, 7, 9, and 10 of AF1. In this analysis the accuracies of responses to the repeated query (i.e., "How many targets will your team successfully photograph in this mission?") were compared. The measure was found to be adequately reliable, as there were no differences among the asymptotic missions, $F(4, 20) < 1, p = .52$.

Next we consider reliability in terms of replicating findings across experiments. The relationship between situation awareness accuracy to the repeated query at Mission 4 and team performance at Mission 4 was observed for each experiment. Situation awareness accuracy significantly predicted team performance in AF1, $F(1, 8) = 4.50, p = .07, \beta = .60$. This positive relationship was replicated in AF3, $F(1, 16) = 3.08, p = .10, \beta = .40$, and also in AF4, $F(1, 18) = 10.86, p < .01, \beta = .61$. Considering all missions, in AF3 and AF4, situation awareness accuracy shows improvement from Missions 1 through 4 (low workload) and a decline in Mission 5 (high workload). These findings suggest that our measure of situation awareness is adequately reliable.

Taskwork knowledge. In Table 106 we report descriptive statistics for the overall accuracy metric for taskwork knowledge across the four experiments which we analyze in this section, as well as for the benchmarking study (AF5).

Table 106
Descriptive Statistics of Taskwork Knowledge Overall Accuracy for each Experiment

Experiment	N	Min	Max	*M*	*SD*	Var
AF 1	41	.36	.58	.48	.06	.004
AF 2	54	.33	.57	.33	.06	.004
AF 3	40	.37	.59	.47	.05	.002
AF 4	20	.35	.59	.47	.06	.003
AF 5	5	.44	.64	.53	.08	.007

N is based on teams x sessions (there were missing data for 3 teams x sessions in Experiment AF1).

The test-retest analysis of the taskwork knowledge accuracy measure rests on a comparison of Sessions 2 and 3 of AF1. For this comparison, none of our taskwork measures showed session-wise changes under $\alpha = .20$. Based on the analyses, we conclude that the taskwork measures have adequate reliability. F values and significance are presented in Table 107.

Table 107
Analyses of Variance for Taskwork Measures

Taskwork Measures	*F*	*p*
Accuracy	.02	.90
Role	.06	.81
IPK	.07	.79
Similarity	.40	.54
Holistic	.11	.75

df = 19

An additional, less formal method of taskwork validation lies in examination of findings that are replicated over those studies that relate to taskwork. Correlations with overall team performance and taskwork for all four experiments revealed that out of all four experiments, only AF1-Session 1 taskwork knowledge measures (all metrics but positional accuracy) were correlated with team performance. It is interesting to note that these data came from a session that was positioned after at least one mission, but not at the end of the experiment. However, the finding does not replicate across experiments.

Another source of validation lies in examination of the shared/non-shared manipulation of AF2 and the co-located/distributed manipulation in AF3 and AF4. Both shared condition teams in AF2, and co-located teams in AF4, had superior taskwork knowledge to non-shared and distributed teams respectively. This pattern of results is in direct contrast to performance results indicating superior performance for non-shared and distributed teams. It seems that these manipulations had an impact on knowledge acquisition in the expected direction, but little impact on team performance. Based on these examinations, we can conclude that the taskwork measures have adequate reliability.

Teamwork knowledge. In Table 108 we report descriptive statistics for the overall accuracy metric for teamwork knowledge across the four experiments we analyze in this section, as well as for the benchmarking study (AF5).

Table 108
Descriptive Statistics of Teamwork Knowledge Overall Accuracy for each Experiment

Experiment	N	Min	Max	*M*	*SD*	Var
AF 1	43	.16	1.00	.64	.21	.05
AF 2	54	.34	.86	.63	.11	.01
AF 3	40	18.33	29.00	24.32	2.28	5.21
AF 4	20	17.00	28.67	23.18	2.73	7.47
AF 5	5	18.00	26.67	23.47	3.58	12.81

N is based on teams x sessions.

No test-retest assessment is possible for teamwork knowledge because the teamwork knowledge measure was different for AF1 and AF2 compared to AF3 through AF5 and there were insufficient session replications in AF3 and AF4. Therefore we assess teamwork reliability in terms of replication of findings across AF3 and AF4.

Across the two experiments there was only an effect of dispersion on the teamwork similarity measure in AF4 and this was not replicated in AF3. There were few correlations of teamwork-associated clusters with performance or process. The only replicated finding included several positive significant correlations between teamwork variables and summary process. Taken together, based on the data that were available to assess the reliability of teamwork knowledge measures suggest weak reliability at best.

Summary. In summary, our reliability assessment indicated adequate reliability for our team performance, process, situation awareness, and taskwork knowledge measures, although data were not available to test our process measures in the test-retest framework. Minimal data were also available to test the reliability of our teamwork measure, because it, like our process summary measure, has evolved over the course of the four experiments. However, based on Experiments AF3 and AF4 it appears that our teamwork measure is weak in terms of reliability.

4.19.2 Validity

The validity of our measures is evaluated in this section in terms of predictive validity and construct validity. Predictive validity is assessed using regression analyses and construct validity is assessed using MTMM matrices. The regression analysis will focus only on AF3 and AF4 because of the similarities in those two studies and because a later analysis involves holistic measures collected consistently only in AF3 (Session 2) and AF4. Also, because we deal with our new holistic measures in two separate sections that follow, these analyses will include only traditional collective measures of team knowledge. The MTMM analysis is only possible with AF3 data because it is in this experiment that we collected secondary measures of taskwork and teamwork. In addition the MTMM analysis focuses only on our taskwork and teamwork measures.

Analysis of predictive validity of measures. This analysis represents an attempt to identify the usefulness of our primary metrics for experiments AF3 and AF4 and both dispersion conditions. Only collective metrics were analyzed and the co-located/distributed distinction is maintained. Each of our metrics will stand on an equal footing with an equal chance to account for team performance variance regardless of any other theoretical considerations.

In order to identify the best possible combination of variables for predicting team performance, exploratory and selection techniques were used on the separate co-located (n = 20) and distributed (n = 20) data. First, all variables (see Table 109) were entered into a linear regression with performance as the dependent variable and dichotomies for workload condition. Note that only one metric per measure was entered into this analysis. Selections were made based on previous results and individual correlations with team performance. Although not shown here, the residuals from this regression were then scatter-plotted by each independent variable. These plots were made in order to identify the need for polynomial terms, to identify unequal variance-

inducing independent variables, and to identify independent variables that might not have independent observations.

The next step performed was Mallow's C_p model selection. This helped us to obtain the best subset of independent variables in terms of explaining performance variance compared to the full set, where "best" means unbiased given our sample of data. More specifically, Mallow's C_p is used to identify an unbiased estimator (as a subset combination of predictors) of model error variance. When the model is unbiased, we expect C_p to be no larger than p, the number of parameters used to estimate mean performance at given predictor levels in the model. Mentally plotting C_p as a function of p, we identified unbiased models as models whose $p \geq C_p$. These models were retained as candidate models for "best in show." Subsequently, candidates were compared in more aesthetic ways. For example, if a model accounts for almost as much performance variance as another, but with fewer variables, then this model was favored.

Table 109
*Variables Used in the Regression Analysis**

- Team Performance = rate of good photos, rate of fuel/film used, et cetera
- Workload = splits performance into two types, one performance score for a team's Mission 4 performance and a second performance score for a team's mission 5 performance
- Critical incident process = ratio of points earned to total points from the following**:

Event	Available Pts.
P1	0-3
P2	0-2
P3	0-1
P4	0-1
P5	0-2
P6	0-1
Total	10

- Taskwork = average of three individual overall taskwork knowledge accuracies within each team
- Teamwork = average of three individual overall teamwork knowledge accuracies within each team
- Situation awareness = sum of individual overall situation awareness accuracy within each team (ranges from 0-3)***

* when quadratic terms are formed for a variable, we label the variable name with a superscript 2

** for missing critical incident process data the available points for that P were subtracted from 10 so as not to penalize the team for following a different route from that which was specified on the score sheet

*** missing situation awareness data replaced with the mean; only repeated situation awareness queries used, non-repeated not sensitive to changes in performance/mission

The best models for co-located and distributed teams for Experiments AF3 and AF4, respectively are presented in Tables 110-113.

Table 110

Experiment AF3 Model for Co-located Teams

Source	*df*	*SS*	*MS*	*F*	*p* > *F*
Model	3	60,368	20,123	14.38	.00
Error	14	19,587	1,399		
Total	17	79,955			

Metric	Estimate	*SE*	*t*	*p* > |*t*|
Intercept	465	59	7.85	.00
Taskwork^2	-899	261	-3.45	.00
SA	31	8	3.95	.00
Critical Incident Process	224	81	2.75	.02

Adj. $R^2 = .703$ $C_p = 3.33$

Table 111

Experiment AF3 Model for Distributed Teams

Source	*df*	*SS*	*MS*	*F*	*p* > *F*
Model	2	32,362	16,181	5.13	.02
Error	17	53,610	3,154		
Total	19	85,972			

Metric	Estimate	*SE*	*t*	*p* > |*t*|
Intercept	804	220	3.66	.00
Workload	-69	25	-2.73	.014
Teamwork	-15	9	-1.67	.11

Adj. $R^2 = .303$ $C_p = 1.68$

Table 112

Experiment AF4 Model for Co-located Teams

Source	*df*	*SS*	*MS*	*F*	*p* > *F*
Model	4	115,913	28,978	7.84	.00
Error	15	55,435	3,695		
Total	19	171,348			

Metric	Estimate	*SE*	*t*	*p* > |*t*|
Intercept	4,211	1,915	2.20	.04
Taskwork	-13,750	7,695	-1.79	.09
Taskwork^2	13,916	7,493	1.86	.08
Teamwork	-21	9	-2.51	.02
SA	41	12	3.49	.00

Adj. $R^2 = .590$ $C_p = 4.14$

Table 113
Experiment AF4 Model for Distributed Teams

Source	*df*	*SS*	*MS*	*F*	*p* > *F*
Model	3	136,989	45,663	24.76	.00
Error	16	29,505	1,844		
Total	19	166,493			

Metric	Estimate	*SE*	*t*	*p* > |*t*|
Intercept	-74	91	-0.82	.43
Taskwork	-1,299	299	4.35	.00
Teamwork^2	-0.31	.11	-2.83	.01
SA^2	17	3	6.21	.00
Adj. $R^2 = .790$	$C_p = 0.88$			

The overlap of variables predictive of team performance across best models is summarized in Table 114. For the co-located subsets, two metrics were in agreement across experiments with taskwork quadratic, however, disagreeing in terms of whether this relationship is concave up (Experiment AF4) or concave down (Experiment AF3). In other words the quadratic metric as applied to Experiment AF4 says middling taskwork scores are the worst, while in Experiment AF3 middling scores are the best with respect to team performance. Nevertheless, this was a useful metric. Part of this discrepancy can be explained by the poor timing of the knowledge sessions in AF3. For this reason, and based on the knowledge data in AF4 that support it, the AF4 finding is probably more reliable. This finding in addition to the significant negative taskwork linear trend also corroborates our general contention that a minimal degree of taskwork knowledge is necessary for good team performance, but better performing teams do not have higher levels of taskwork knowledge. That is too little taskwork knowledge is not sufficient for good team performance; too much taskwork knowledge may mean that time was ill spent on acquiring factual knowledge at the cost of team process skill. Situation awareness was the other measure that was important across the co-located subsets. In both of these cases, high situation awareness was related to high team performance.

For distributed subsets, only collective teamwork agreed. It should be noted however that none of the collective metrics were really adequate for the distributed teams in Experiment AF3 (i.e., at $\alpha = .10$). Additionally the linear trend was not significant for distributed teams in Experiment AF4, so consistency is further limited. Inspecting Table 114 down the columns there was complete agreement among subsets for Experiment AF4, but none for Experiment AF3. Again, this may be due to the poor placement of the knowledge sessions in AF3. Across all cells, taskwork and collective situation awareness agreed in three out of four cells as important predictors. Similarly, teamwork occurred in three cells out of four, although as noted before this is fairly unreliable given that it was only marginally successful as a predictor for distributed teams in Experiment AF3. Interestingly, critical incident process was not as good as a predictor of team performance as the team knowledge variables.

Table 114
Agreement of Significant Factors in Experiment AF3 and Experiment AF4 Co-located and Distributed

	Experiment AF3	Experiment AF4	Number agree
Co-located	Taskwork Situation Awareness Critical incident process	Taskwork Teamwork Situation Awareness	2
Distributed	Workload Teamwork	Taskwork Teamwork Situation Awareness	1
Number agree	0	3	across cells = 0

Finally, we looked at the predictor subsets and identified the metric in each with the highest partial correlation with team performance. Table 115 lists these by subset. Overwhelmingly, situation awareness was a good metric in terms of performance across subsets. The fact that taskwork and teamwork knowledge accounted for less performance variance than team situation awareness may speak not only to the poor placement of the knowledge sessions in AF3, but also to the use of the traditional collective metrics that may not be appropriate for characterizing the knowledge of heterogeneous teams. This issue will be explored more fully in the later sections that discuss collective vs. holistic metrics.

Table 115
Metric With Highest Partial Team Performance Correlation for Each Subset

	Condition	
	Co-located	Distributed
Experiment AF3	Situation Awareness	Teamwork
Experiment AF4	Situation Awareness	Situation Awareness

MTMM. This analysis was conducted on data from Experiment AF3 in order to assess construct validity of taskwork and teamwork knowledge measures. First we provide a brief overview of this method. MTMM matrices (Campbell & Fiske, 1959) are used to assess construct validity for two or more constructs measured by two or more sources. Construct validity contains two sub-categories: (1) convergent validity and (2) divergent validity. Convergent validity refers to the relatedness of two different measures of the same theoretical construct. Relatedness of two measures of the same construct should be high for convergent validity. Divergent validity refers to the relatedness of two measures of different constructs. Relatedness of two different constructs should be low for divergent validity. When *both* convergent and divergent validity are evidenced, construct validity is supported among the constructs being assessed. The MTMM matrix contains information which addresses both convergent and divergent validity between two or more constructs and thus assesses construct validity. The MTMM matrix is simply a correlation matrix arranged to facilitate assessment of construct validity. The requirements for the proper arrangement are multiple traits nested within multiple methods. That is, one method

is used to measure multiple constructs; a separate method is then used to measure each of the multiple constructs, etc. Obviously, all constructs have to be measured by each method for a full MTMM analysis.

After sorting the correlations into the proper format, the MTMM matrix can then be broken down into component sub-matrices that tell the tale of validity (see Figure 36).

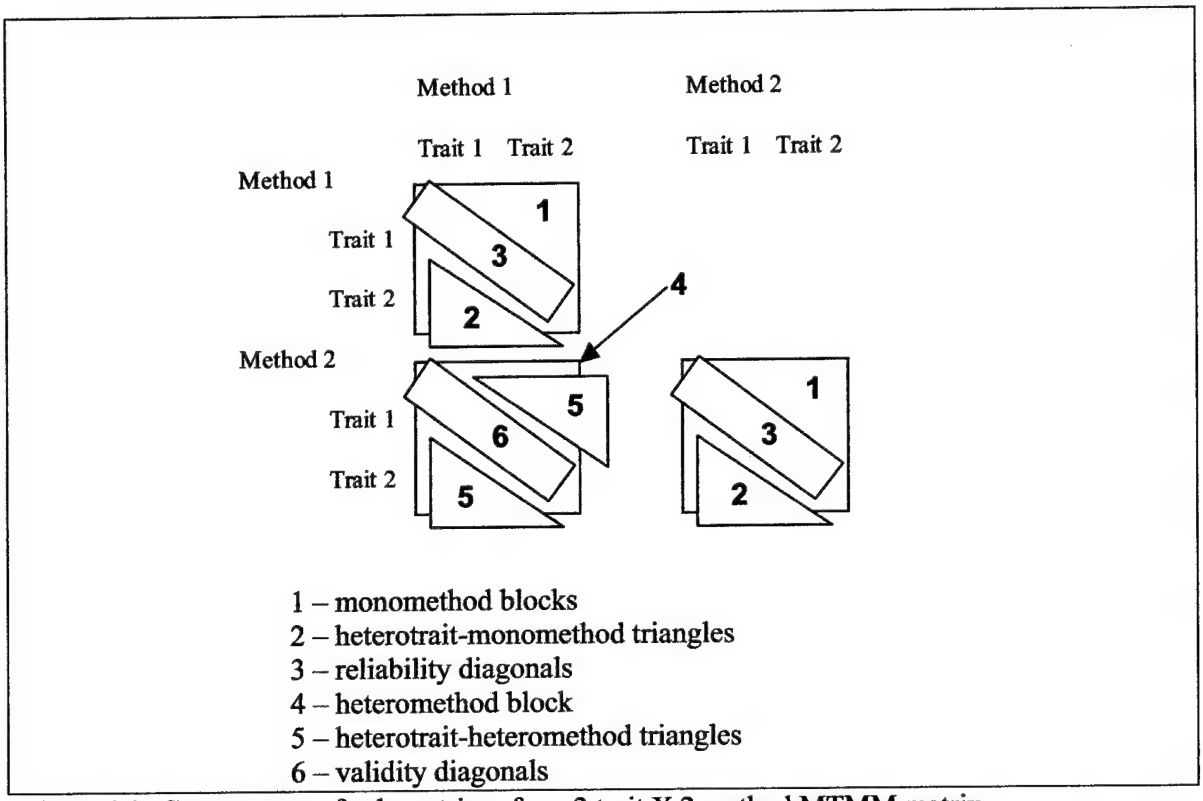


Figure 36. Components of sub-matrices for a 2 trait X 2 method MTMM matrix.

Monomethod blocks are associated with the methods used for measuring the constructs. There are as many monomethod blocks as measurement methods. Monomethod blocks contain the heterotrait-monomethod triangles. High heterotrait-monomethod correlations indicate that the method of measurement leads to correlated measures on different constructs. This result can be interpreted as a strong measurement effect. Negative or low heterotrait-monomethod correlations are evidence of divergent validity. The monomethod blocks also contain the reliability diagonals, where estimates of reliability across multiple samples can be placed. Heteromethod blocks contain information about the constructs when measured by different methods. If m is the number of methods used, there are $(m*m\text{-}1) / 2$ of these. The heteromethod blocks contain the heterotrait-heteromethod triangles and the validity diagonals. The heterotrait-heteromethod triangles contain correlations that share neither trait nor method. Small and negative correlations here are evidence of divergent validity. The validity diagonals are also in the heteromethod blocks. High positive correlations here indicate convergent validity. As a rule of thumb, convergent validity is evidenced when this correlation is larger than all other correlations in its associated row and column of their heterotrait-heteromethod triangles.

Additionally, these correlations should be larger than the correlations in the heterotrait-monomethod triangles. This result suggests that trait effect is stronger than the measurement effect.

We used two different methods for measuring the taskwork and teamwork knowledge constructs as is dictated by the MTMM paradigm. Knowledge measures taken in the second session were used for the MTMM analysis since this knowledge is presumably more stable. Our methods included our standard taskwork and teamwork knowledge measures and a multiple-choice test of taskwork and teamwork knowledge (see secondary knowledge questions in Appendix J). Individuals received a percent correct on each of the two multiple-choice tests (i.e., taskwork and teamwork). Scores were averaged across the three team-members in a team to get a team aggregate score.

Our standard taskwork measure involves pairwise relatedness ratings of task-relevant concepts, which are submitted to Pathfinder network scaling. Knowledge scores are based on the similarity of the resulting network to a referent. Teamwork knowledge is assessed in a questionnaire in which individuals check the critical information needed and sender and receiver in a given scenario. Teamwork knowledge is also scored by comparison of the checked responses to a referent. For this analysis, the overall accuracy scores of taskwork and teamwork were used, where the overall accuracy knowledge scores were averaged across team members to generate a collective score.

A MTMM matrix was constructed for each team member and the team overall. Scores on multiple-choice tests were proportion correct out of the total possible (= x/5). These were correlated with overall teamwork and taskwork. The correlations in the matrices are based on 20 observations each except for correlations with taskwork knowledge ratings. Due to missing data (Team 7) these correlations involve only 19 pairs of observations.

Table 116 presents the MTMM matrix for AVOs. The small negative correlations in the heterotrait-monomethod correlations provide evidence of divergent validity. The positive value in the validity diagonal for teamwork knowledge provides some evidence for convergent validity for this construct. This correlation for teamwork knowledge is evidence of a larger trait effect than measuring method effect for teamwork knowledge.

Table 116
AVO Taskwork and Teamwork Knowledge Correlation Matrix

	Multiple Choice		Standard Method	
	Task	Team	Task	Team
Mult. Choice Taskwork	1.00			
Mult. Choice Teamwork	-.07	1.00		
Standard Taskwork	-.10	.03	1.00	
Standard Teamwork	.11	.34	-.03	1.00

The MTMM matrix for PLOs is presented in Table 117. There is good evidence for divergent validity in the heterotrait-monomethod correlations, but no evidence for convergent validity in the validity diagonals. Construct validity for either construct is not supported in the PLO results. An opposite pattern than expected emerged. Taskwork and teamwork were positively correlated when using different methods of measurement, but not when using the same.

Table 117
PLO Taskwork and Teamwork Knowledge Correlation Matrix

	Multiple Choice		Standard Method	
	Task	Team	Task	Team
Mult. Choice Taskwork	1.00			
Mult. Choice Teamwork	-.37	1.00		
Standard Taskwork	-.25	.42*	1.00	
Standard Teamwork	.27	.07	-.10	1.00

* $p < .10$

The DEMPCs' MTMM matrix is given in Table 118. As with the PLOs, the DEMPCs MTMM demonstrated divergent validity in the heterotrait-monomethod diagonals. DEMPC divergent validity was especially apparent in the ratings. There is no evidence for convergent validity in the validity diagonals. As with the PLO MTMM, the high teamwork multiple-choice to taskwork ratings correlation is completely opposite of that expected. This result seemingly contradicts the evidence of divergent validity found in the heterotrait-monomethod diagonals.

Table 118
DEMPC Taskwork and Teamwork Knowledge Correlation Matrix

	Multiple Choice		Standard Method	
	Task	Team	Task	Team
Mult. Choice Taskwork	1.00			
Mult. Choice Teamwork	-.07	1.00		
Standard Taskwork	.02	.50*	1.00	
Standard Teamwork	-.12	-.19	-.35	1.00

* $p < .05$

For the team level MTMM, ratings accuracy, averaged across the three team members, was correlated with multiple-choice scores which were also averaged across the three team members. The overall team MTMM is given in Table 119. These two constructs at the team level show no evidence of convergent validity (refer to the validity diagonal in Table 119). Furthermore, the correlation between multiple-choice taskwork and teamwork ratings is the largest in the matrix.

This suggests that not only there is no evidence of convergent validity, but there is little here to support these constructs' divergent validity either.

Table 119
Team Taskwork and Teamwork Knowledge Correlation Matrix

	Multiple Choice		Standard Method	
	Task	Team	Task	Team
Mult. Choice Taskwork	1.00			
Mult. Choice Teamwork	-.10	1.00		
Standard Taskwork	-.08	.17	1.00	
Standard Teamwork	.24	-.06	-.16	1.00

Construct validity is supported when two features are evidenced in a MTMM analysis: (1) convergent validity and (2) divergent validity. Construct validity was supported only for AVO teamwork knowledge. Teamwork knowledge perhaps became more intermingled with taskwork knowledge for PLOs, DEMPCs, and at the team level. Is it something about the AVOs' duties that kept teamwork knowledge isolated? Teamwork knowledge, as implied above, involves knowing what others know, what others need to know, and how to get information to the proper team member. AVO, in a sense, can become a conduit between DEMPC and PLO over missions. In a rough analysis, the AVO collects specific information from DEMPC, then based on this information, flies the UAV within a certain proximity to a target, at which point the AVO coordinates with PLO in order for the team to accomplish their most important task, taking a picture. By repeating this scenario numerous times over seven missions, the AVO may develop the most well defined sense of who knows what, who needs to know what, and when and how information needs to be delivered in order for the team to succeed. It would not be surprising, given this premise, that a test of AVO knowledge would provide support for a valid psychological construct involving exclusively teamwork knowledge.

One limitation of this approach that should be acknowledged is the reliance on linear relationships evident in the correlations. It is possible that nonlinear relationships exist between these factors. In addition the poor placement of the knowledge sessions in Experiment AF3 may contribute to measurement error. In addition, the small sample of 20 Experiment AF3 teams further limits the statistical power available for this analysis.

Summary. In sum, the traditional collectively-oriented knowledge measures used in these analyses did not demonstrate exceptional predictive or construct validity. Our situation awareness measure and, to a lesser extent, our taskwork measure were predictive of team performance. However, as discussed in previous sections, we are not convinced that the situation awareness measure reflects what we mean by situation awareness. However, regardless of the source of accurate responses on the repeated situation awareness queries, they are predictive of performance. Instead of a measure of situation awareness, we view the queries as tests of test-taking skills. To score well on the repeated queries, teams have to appreciate what is

important in this task and basically learn how to play the game. These good "guessers" also tend to be good team members. Whether they also have good situation awareness as a team in a dynamic environment is an open question.

The low validity of our taskwork and teamwork measures both in terms of our predictive validity analysis and MTMM can be explained in a number of ways including the poor knowledge session placement. In addition, it may be that the collective measures are indeed inadequate for capturing team knowledge of a heterogeneous team. Further, it may be that declarative knowledge is not predictive of team performance past a certain point. One or more of these explanations may hold. In the next two sections, we further explore the collective vs. holistic measurement issue.

4.19.3 Collective vs. Holistic Measures

In this section we explore in depth the concept of holistic assessment of a team. Traditionally, team measures are taken at the individual level and then aggregated. Alternatively, our holistic measures are taken at the team level and include our team performance measure, which is based on a composite of outcome measures relevant to the team's goals, and our consensus knowledge measures in which teams come to consensus on a response to a knowledge probe.

Before comparing our collective and holistic measures we will say more about aggregation schemes associated with collective measures. Collective measures have two limitations in regard to team measurement. First, collective measures tend to overlook team interaction or team process. The aggregation scheme tends to serve as a model of team process. However, most commonly overly simplistic aggregation schemes (such as averaging) are applied for measuring team process, thereby, limiting applicability of the measure to heterogeneous teams in which all team members are not equal.

Generally, the collective level consists of taking an arithmetic mean of individual scores on the measure in question. However, other aggregation schemes may be appropriate as well, depending on how teams interact to produce the final team score (Steiner, 1972). For instance, if teams tend to follow their best performer in coming to team decisions, then the maximum is a better aggregate than the arithmetic mean. If the task is such that the worst performer limits team performance, then the minimum is more appropriate. The geometric mean is an aggregate that combines the features of the minimum and the arithmetic mean. The geometric mean is the product of the individual cases, taken to the n^{th} root.

This issue becomes especially important when considering measures that are not available at a holistic level. There is no obvious way to create a holistic score for individual difference measures, such as working memory, GPA, the TLX workload measure, and processing speed. For such individual measures, it is not clear what a team score should be, or even if there should be a team score. With repeated measures and multivariate techniques, it is possible to treat the vector of individual scores as a vector, with no attempt to integrate. This greatly adds to the complexity of the models, however. Moreover, summary measures based on samples drawn from the same population are more stable parameter estimates than individual observations. For

individual-level measures, it is generally preferable to create a team-level aggregate for these reasons, that is, if it is conceptually possible to create a team score.

In the context of our UAV-STE we use averaging or summing to generate our collective metrics, however it would be more appropriate to rely on individual roles. We see role effects with many of our individual difference measures. In that case, an appropriate aggregate might be to take a weighted average. For instance, knowing that GPA is more correlated with performance for the PLO than for the other two team members, we can reflect the interdependence among team members by somehow giving PLO a higher weight than other team members. If a measure is conceptually driven so much by the PLO that other team members are not relevant, then the team aggregate would be the PLO's individual score. Similarly, if task characteristics are such that the team totally relies on DEMPC's working memory, and no-one else's, then the aggregate would be DEMPC's individual score. The aggregate that comes closest to simulating a holistic score, will be whatever aggregate (1) best reflects the interdependence among team members, and (2) accounts for the characteristics of the task.

In the following two sections we examine our data in light of the collective vs. holistic distinction. The first question that we ask is does this distinction matter? Do collective and holistic team performance scores result in differential rankings of teams? The following section will explore the relative predictive validity of collective vs. holistic measures of knowledge.

Does it make a difference? To explore issues of collective versus holistic measurement in greater depth, we compared our usual holistic measure of performance with a collective measure, to determine whether or not the two measurement approaches would rank order teams similarly. The collective measure was defined as the average of the three individual performance scores, which were first standardized within each mission for each of the four studies. Three measures of concurrence were used. One of them was the Spearman rank order correlation coefficient (R_s). We also examined the top five performing teams as defined by the holistic measure, compared with those defined by the collective measure. We computed the *C-value* between these two sets, defined as the intersection divided by the union. At last, we computed the *C-value* between the two sets of the five lowest-performing teams.

Table 120 shows the Spearman, highest five C-value, and lowest five C-value measures for each mission within each study. C-values tend to be below .5 for the high ranking teams, and above .5 for the low ranking teams. This indicates that there is more consistency in ranking which teams did poorly, than in ranking those that did well. Spearman correlations range from 0 to .92, but tend to hover around .5. As seen in Figure 37, a consistent growth pattern of Spearman correlations emerges across missions. The correlations show a sporadic increase across missions, peaking at approximately Mission 4. This is also approximately where performance asymptotes. In AF1, the Spearman correlation even drops at Mission 8, when performance drops.

Table 120

Rank Order Concurrence Between Collective and Holistic Performance Measures

Exp. AF	Mission	Cases	Spearman	High 5 C	Low 5 C
1	1	11	0.12	0.43	0.43
1	2	11	0.36	0.43	0.67
1	3	11	0.93	0.67	1.00
1	4	11	0.75	0.67	0.67
1	5	11	0.67	0.43	0.67
1	6	11	0.83	0.43	0.67
1	7	11	0.58	0.43	0.67
1	8	10	0.32	0.43	0.43
1	9	10	0.77	0.67	0.67
1	10	9	0.77	0.67	0.67
2	1	18	-0.001	0.25	0.25
2	2	18	0.60	0.43	0.43
2	3	18	0.45	0.25	0.25
2	4	18	0.76	0.43	0.67
2	5	18	0.48	0.43	0.43
3	1	20	0.16	0.11	0.67
3	2	20	0.52	0.67	0.43
3	3	20	0.45	0.11	0.67
3	4	20	0.52	0.43	0.67
3	5	20	0.61	0.43	0.67
3	6	20	0.75	0.43	0.43
3	7	20	0.41	0.11	0.67
4	1	20	0.44	0.43	0.25
4	2	20	0.80	0.43	0.67
4	3	20	0.72	0.43	0.67
4	4	20	0.90	0.67	1.00
4	5	20	0.72	0.43	0.67

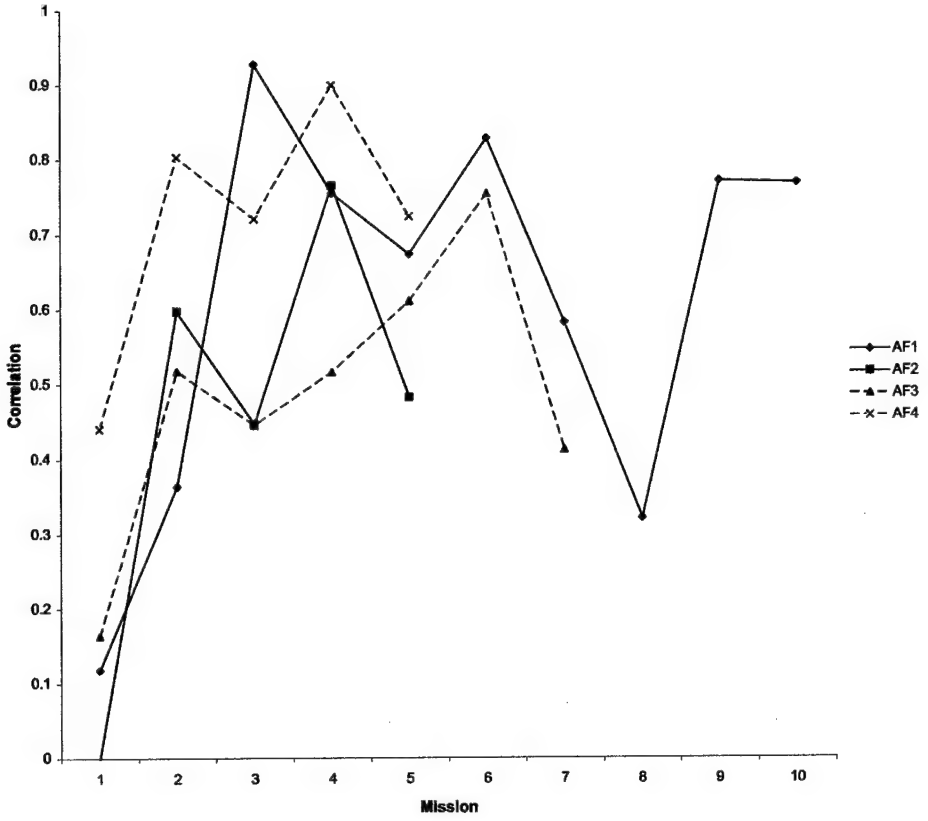


Figure 37. Growth pattern for Spearman correlations for all studies.

In addition to these three measures, we identified teams who repeatedly crossed over from the highest five cases on one measure, to the lowest five cases on the other. Since these teams show repeated instances of high rank on one measure, but low rank on the other, it is possible that they can be used to identify more specifically what interaction patterns define the differences between holistic and collective measures.

Teams who repeatedly crossed from high to low between collective and holistic measures were common in the first study, because there were only 9, 10, or 11 cases. Data are not considered for the missions in which there were only 9 valid cases, because at least one team must cross over in such a situation. In AF1, Teams 1 and 8 crossed over in four and five of the nine missions, respectively. These teams were split approximately evenly between high collective-low holistic combinations, and the reverse. Teams 2 and 3 showed, respectively, two and three instances of high-holistic/low-collective ranks. Teams 6 and 7 each showed two instances of high-collective/low-holistic crosses.

The other three studies showed considerably fewer teams with repeated cross-overs of rank, between measures. In AF2, Team 9 had two instances of a high-collective/low-holistic rank combination, as did Team 3 in AF3. AF4 showed no repeated patterns.

In conclusion, it is apparent from these rank order findings that collective and holistic measures of performance do yield different results. Ranking agreement was lowest for the ranking of the five best teams, and overall correlations were moderate. However, the discrepancy seems to dissipate over time, as performance itself increases. This may be due to decreases in variance, or increases in seamless team interaction patterns. In any event, aggregation schemes become less relevant as teams mature, because holistic and collective measures converge. Individual performance scores become more closely intertwined, because they come closer to resembling the holistic team performance score. In future research, more specific differences between the measurement approaches can be examined further in light of the specific teams who showed marked differences between the measures.

Does the difference matter? Although we have detected different outcomes dependent on whether collective or holistic methods are applied, we have not determined whether one is more valid than the other. In this section we compare our knowledge measures (situation awareness, taskwork, and teamwork) measured collectively or holistically. This analysis was conducted identically to the previous analysis in the validity section. In this previous analysis collective measures of situation awareness, taskwork knowledge, and teamwork knowledge were entered into a regression equation with critical incident process to predict team performance. Results revealed that situation awareness and, to a lesser extent, teamwork knowledge were the best predictors of team performance. In this section, the analysis will be repeated, but this time including holistic measures of situation awareness, taskwork knowledge, and teamwork knowledge (see Table 121).

Table 121
Holistic Variables Included in the Regression Analysis.*

- Holistic taskwork knowledge = team (consensus) network compared to team referent
- Holistic teamwork knowledge = team's (consensus) overall accuracy
- Holistic situation awareness accuracy= 1 or 0, is team accurate (1) or not (0)**

* when quadratic terms are formed for a variable, we label the variable name with a superscript 2

** missing situation awareness data replaced with the mean; only repeated situation awareness queries used, non-repeated not sensitive to changes in performance/mission

The best models for co-located and distributed teams for Experiments AF3 and AF4, respectively are presented in Tables 122-125. Knowledge measures are preceded by the term "collective" or "holistic" in these tables to distinguish them.

Table 122

Experiment AF3 Model for Co-located Teams

Source	*df*	*SS*	*MS*	*F*	*p* > *F*
Model	5	67,238	13,448	18.05	.00
Error	10	7,451	745		
Total	15	74,689			

Metric	Estimate	*SE*	*t*	*p* > |*t*|
Intercept	184	128	1.44	.18
Collective Taskwork^2	-832	216	-3.94	.00
Critical Incident Process	168	64	2.63	.03
Holistic Taskwork^2	291	90	3.23	.01
Holistic Teamwork	7	4	1.82	.10
Holistic SA	94	17	5.66	.00
Adj. R^2 = .850	C_p = 3.65			

Table 123

Experiment AF3 Model for Distributed Teams

Source	*df*	*SS*	*MS*	*F*	*p* > *F*
Model	4	55,758	13,939	6.92	.00
Error	15	30,214	2,014		
Total	19	85,972			

Metric	Estimate	*SE*	*t*	*p* > |*t*|
Intercept	874	178	4.92	.00
Collective Taskwork^2	-1008	287	-3.51	.00
Holistic Taskwork^2	369	152	2.42	.03
Holistic Teamwork	-15	5	-2.81	.01
Holistic SA	87	22	4.00	.00
Adj. R^2 = .555	C_p = 2.14			

Table 124
Experiment AF4 Model for Co-located Teams

Source	*df*	*SS*	*MS*	*F*	*p* > *F*
Model	5	130,829	26,166	9.40	.00
Error	14	40,519	2,894		
Total	19	171,348			

Metric	Estimate	*SE*	*t*	*p* > |*t*|
Intercept	6,577	1,736	3.79	.00
Collective Taskwork	-28,235	6,780	-4.16	.00
CollectiveTaskwork^2	27,259	6,545	4.16	.00
Holistic Teamwork	102	42	2.38	.03
Holistic Teamwork^2	-2	1	-2.59	.02
Collective SA	31	11	2.94	.01

Adj. $R^2 = .679$ $C_p = 4.85$

Table 125
Experiment AF4 Model for Distributed Teams

Source	*df*	*SS*	*MS*	*F*	*p* > *F*
Model	7	154,662	22,095	22.41	.00
Error	12	11,831	986		
Total	19	166,493			

Metric	Estimate	*SE*	*t*	*p* > |*t*|
Intercept	1,022	166	6.17	.00
Workload	-109	30	-3.60	.00
Collective Taskwork^2	3,159	482	6.56	.00
Holistic Teamwork	-30	6	-5.48	.00
Collective SA^2	14	4	3.31	.01
Holistic SA	-88	29	-3.02	.01
Critical Incident Process	-1,773	439	-4.03	.00
Critical Incident Process^2	1,595	413	3.86	.00

Adj. $R^2 = .888$ $C_p = 8.95$

On the positive side, we were able to find combinations of metrics in each data set that were significantly better at accounting for performance variance than the null model. As can be seen in Table 126, the metrics for Experiment AF3 largely agreed, although the directionality of the holistic teamwork estimate was positive for co-located and negative for distributed. A similar pattern can be seen in Experiment AF4, with holistic teamwork being important in both, but with the effect being positive for co-located and negative for distributed, which agrees with the Experiment AF3 findings.

Next, looking across the rows of Table 126, there is not much agreement among the subsets in terms of condition (i.e., co-located vs. distributed) across experiments, although there was some. For co-located two metrics were common among the subsets, collective taskwork^2 and holistic teamwork, however only holistic teamwork agreed in sign (positive). Collective taskwork^2 had a negative weight for AF3 co-located teams and a positive weight for AF4 co-located teams. These two were also common among the distributed metrics in addition to holistic situation awareness. The distributed analyses agreed on directionality for holistic teamwork (negative), and although they disagreed in terms of direction on holistic situation awareness (Experiment AF3 was positive; Experiment AF4 was negative) and collective taskwork^2 which was positive for AF4 and negative for AF3.

Next, examining Table 126 across the cells, we can see that two of the metrics were common among all of the subsets: collective taskwork^2 and holistic teamwork. Aside from being significant global predictors in each analysis, the estimates for these metrics suggest that the direction of the relationship for collective taskwork^2 depends on experiment. Interestingly, the estimates for holistic teamwork indicate that holistic teamwork shares a partial relationship with team performance across experiments in which high holistic teamwork was related to high performance for co-located teams, but low performance for distributed teams.

Table 126
Agreement of Significant Global Subsets in Experiment 1 and Experiment 2 Co-located and Distributed

	Experiment AF3	Experiment AF4	Number agree
Co-located	Collective taskwork^2 Holistic taskwork^2 Holistic teamwork Holistic SA Critical incident process	Collective taskwork^2 Holistic teamwork Collective SA Collective taskwork Holistic teamwork^2 Collective taskwork^2	2
Distributed	Collective taskwork^2 Holistic taskwork^2 Holistic teamwork Holistic SA	Holistic teamwork Collective SA^2 Holistic SA Critical incident process Critical incident process ^{2}	3
Number agree	4	2	across cells = 2

SA = Situation awareness

As before, we looked at the predictor subsets and identified the metric in each with the highest partial correlation with team performance. Table 127 lists these by analyses. Similar to the previous analysis with collective metrics only, situation awareness was a good metric. When holistic metrics were included, the holistic situation awareness measure had the highest partial correlation in Experiment AF3.

Table 127
Metric With Highest Partial Team Performance Correlation for Each Subset

	All metrics	
	Co-located	Distributed
Experiment AF3	Holistic SA	Holistic SA
Experiment AF4	Collective taskwork	Collective taskwork

In general, the amount of global team performance variance accounted for (Adj. R^2) increased with the addition of the holistic metrics from .30 to .79 for the models without holistic metrics and from .56 to .89 for the models with holistic metrics. This pattern suggests that holistic metrics are useful for capturing important aspects of team performance. Agreement over all analyses also increased with the inclusion of holistic metrics. When we included the holistic metrics, collective taskwork[2] and holistic teamwork were included in the best predictor subsets over all conditions, exhibiting some consensus of what is important in terms of team performance from a global perspective. As noted above, holistic teamwork was positive for co-located and negative for distributed teams. This result inspires a belief that holistic metrics may additionally be sensitive to differences in team distribution, differences which may not accrue in the static aggregate knowledge of team members, but rather by *how* that knowledge is combined through coordination.

Summary. We have shown that in terms of team performance, collective versus holistic measurement makes a difference in the rank ordering of teams. Further we have shown that holistic knowledge measures account for variance in team performance not accounted for by traditional collective metrics. Further, there is some indication that the way in which holistic metrics relate to team performance may be sensitive to differences in team dispersion.

For now we conclude that collective and holistic measures result in different outcomes and that both forms of measurement serve a purpose in accounting for variance in team performance. We additionally speculate that whereas the collective measures of teams may provide a good representation of the aggregate knowledge of team members, the holistic metrics may do a better job at representing cognitive processing that occurs at the team level. To the extent that this processing is important and not captured by the collective aggregation scheme, holistic metrics should be useful.

4.19.4 Inferring Team Process from Holistic Decision Strategies

We assume that holistic measures of knowledge reflect team knowledge that has been processed by the team through the interactions that take place during the consensus task. The holistic measure should be a good measure of team knowledge to the extent that the process used by the team in the consensus measurement task maps onto that of the actual task – in this case the UAV-STE missions. It is possible, in fact, through a comparison of the individual (i.e., collective) measures (without process) to the holistic measures (with process) to infer the process that was used by a team. In this section we compare individual to collective responses to identify

the decision scheme that was used by the team in coming to consensus. We then determine whether the decision schemes identified are useful in predicting team performance.

For each measure in this analysis and each individual knowledge probe (i.e., situation awareness query, taskwork rating, teamwork judgment) we examine the three individual responses by each of the three team members as well as the consensual team response. Each set of four responses was then classified according to one of six rules that mapped individual responses onto the team response (see Table 128). A SAS (Statistical Analysis Software) program was developed to determine the proportion of instances in which each of the teams used the different strategies.

Table 128
Categories of Responses Made by Three Individuals and the Team

1) Unanimous: team response agrees with all three individuals (e.g., AVO=1, PLO = 1, DEMPC = 1, Team = 1)
2) Majority rules: team response agrees with 2 out of 3 team members (e.g., AVO = 1, PLO=1, DEMPC = 0, Team =1)
3) None: team response agrees with no individual member (e.g., AVO=1, PLO=1, DEMPC=1, Team=0)
4) Leader: team response agrees with only one team member (e.g., AVO=1, PLO=0, DEMPC=0, Team =1). This strategy was further broken down to reveal which individual's rating agreed with the team rating (Leader = AVO + PLO + DEMPC)
    a. Team response agrees with AVO only
    b. Team response agrees with PLO only
    c. Team response agrees with DEMPC only
5) Middle: no one agrees with team rating but team rating is in the mid range of the other ratings (e.g., AVO=5, PLO=6, DEMPC=9, Team =7)
6) Average: team response is an average of the individual responses

Situation awareness. As described in methods section, situation awareness was measured at the individual and team levels by administering queries to each individual and then to the team as a whole, requiring the team to reach a consensus. In order to identify strategies that the teams used to come to consensus in responding to situation awareness queries, the three individual responses and one team response to each query were examined for each of the twenty teams. Recall that at each of the seven missions, teams responded to a repeated query and a non-repeated query for a total of seven repeated queries and seven non-repeated queries per team.

Analyses were conducted separately for the repeated and non-repeated queries. The correct answer to the repeated query, which asked for a prediction about the number of targets the team would photograph by the end of a mission, was not available on any team member's display and therefore required a long-term prediction. In low workload missions responses ranged from 0 to 9, and in the high workload missions there were 20 targets possible resulting in a range of responses from 0 to 20. Mathematical strategies, such as averaging responses, could be used in reaching consensus on the repeated query. However, only non-mathematical strategies (e.g., majority rules) could be used to reach consensus on the non-repeated queries. These queries

asked for the names of targets, types of waypoints, etc. It would have been possible to use mathematical strategies to reach consensus on two of the non-repeated queries, that asked for a short-term prediction on what airspeed or altitude would be. However, upcoming airspeed and altitude are known by the AVO; therefore, we thought it would be very unlikely that a team would decide airspeed or altitude by averaging their individual responses. Therefore only the first four categories listed in Table 128 were applied to the non-repeated queries. All six categories were applied to the repeated queries. Finally, the analyses did not take workload into consideration. By breaking up the analyses by workload, the proportions of time teams used each strategy would be based on so few missions (i.e., 4 missions in low workload and 3 missions in high workload). Missing data would further decrease the sample size, making proportions meaningless.

Across all teams in Experiment AF3, there were missing non-repeated query data for 25 missions where the individual responses could not be compared to team responses because the situation in question changed during the course of administering the question to all individuals and the team. As Table 129 shows, many teams relied upon multiple strategies equally, as indicated by equal proportions. For cases in which the leader strategy was used, such that the team went with one team member's response, the light gray shading indicates which team member's response was typically chosen as the team's response.

It appears that most co-located teams used the unanimous and leader strategies. That is, they either went with their unanimous individual responses or with the single individual who claimed to have knowledge in the area. When teams relied on the leader strategy, it was usually the DEMPC's response that was given as the team response. Distributed teams for the most part used the unanimous strategy followed by the leader strategy. When the leader strategy was used, it was typically the DEMPCs response that was chosen.

Overall, a two-way chi square analysis indicated that the proportion of times each strategy was used did not depend on whether teams were co-located or distributed, $\chi^2(3) = .46$.

Multiple regression analyses assessed whether the strategies the teams used in reaching their holistic response to the situation awareness queries were predictive of their performance. The pattern of strategies used failed to significantly predict performance, $F(4, 15) < 1$.

Table 129

*Mapping Individual Responses to Team Responses on the Non-repeated Situation Awareness Queries for Co-located and Distributed Teams of Experiment AF3**

Team	Unanimous	Majority	None	Leader	AVO	PLO	DEM	N
Co-located Teams								
1		0.17	0.17		0.00	0.00	0.33	6
2		0.14	0.00	0.14	0.14	0.00	0.00	7
3		0.17	0.17		0.17	0.00	0.17	6
7	0.17		0.17		0.17	0.00	0.17	6
8		0.00	0.00		0.17	0.00	0.33	6
11	0.17	0.17			0.17	0.00	0.17	6
12	0.17	0.17	0.17		0.00	0.00	0.50	6
13		0.29	0.00	0.14	0.00	0.00	0.14	7
14		0.14	0.00	0.43	0.29	0.00	0.14	7
16		0.00	0.00	0.20	0.00	0.00	0.20	5
Distributed Teams								
4		0.40	0.00	0.00	0.00	0.00	0.00	5
5		0.00			0.17	0.00	0.17	6
6	0.17	0.17	0.17		0.00	0.00	0.50	6
9		0.17	0.00	0.17	0.00	0.00	0.17	6
10		0.20	0.00	0.20	0.00	0.00	0.20	5
15		0.00	0.00	0.40	0.40	0.00	0.00	5
17		0.25	0.00	0.25	0.00	0.00	0.25	4
19		0.14	0.29	0.14	0.14	0.00	0.00	7
20	0.20		0.00		0.00	0.20	0.20	5
21	0.00	0.25	0.00		0.25	0.25	0.25	4

*Highlighted cells indicate which strategies teams used most.

Now repeating this analysis on non-repeated queries for Experiment AF4, across all teams, there are missing data for 19 missions where the individual responses could not be compared to team responses because the situation in question changed during the course of administering the question to all individuals and the team. As Table 130 shows, many teams relied upon multiple strategies equally, as indicated by equal proportions. It appears that for the most part co-located teams used the majority rules strategy, followed by the leader strategy, and finally the unanimous strategy. When co-located teams relied on the leader strategy, they always chose the DEMPC's response as the team response. Distributed teams relied upon the majority rules strategy to a lesser extent than the co-located teams. Instead, distributed teams tended to reach consensus by using the unanimous strategy or the leader strategy. When the leader strategy was used and a single team member's response was given as the team response, it was almost always the DEMPC's response that was chosen.

Table 130

*Mapping Individual Responses to Team Responses on the Non-repeated Situation Awareness Queries for Co-located and Distributed Teams of Experiment AF4**

Team	Unanimous	Majority	None	Leader	AVO	PLO	DEM	N
Co-located Teams								
4	0.25		0.00	0.25	0.00	0.00	0.25	4
5	0.00				0.00	0.00	0.33	3
6	0.00				0.00	0.00	0.33	3
8		0.00	0.25	0.25	0.00	0.00	0.25	4
9	0.00		0.00	0.20	0.00	0.00	0.20	5
10	0.20	0.20	0.20		0.00	0.00	0.40	5
12		0.33	0.00	0.00	0.00	0.00	0.00	3
15	0.20		0.00	0.20	0.00	0.00	0.20	5
16	0.00		0.00	0.00	0.00	0.00	0.00	3
18	0.00		0.20		0.00	0.00	0.40	5
Distributed Teams								
1	0.25	0.00	0.25		0.00	0.00	0.50	4
2		0.00	0.25	0.25	0.00	0.00	0.25	4
3	0.25		0.00	0.25	0.00	0.00	0.25	4
7		0.00	0.00	0.33	0.00	0.00	0.33	3
11	0.25	0.25	0.00		0.00	0.00	0.50	4
13	0.20		0.00		0.20	0.00	0.20	5
14		0.00	0.00		0.00	0.00	0.50	4
17		0.00	0.00	0.25	0.25	0.00	0.00	4
19		0.20	0.00		0.20	0.00	0.20	5
20	0.25		0.00	0.25	0.25	0.00	0.00	4

*Highlighted cells indicate which strategies teams used most.

Overall, a two-way chi square analysis indicated that the proportion of times each strategy occurred depended on whether teams were co-located or distributed, $\chi^2(3) = 9.64$, $p < .05$. *Post hoc* one-way chi square tests revealed that co-located and distributed teams use the unanimous strategy and majority rules strategy differently, $\chi^2(1) = 3.52$, $p < .10$ and $\chi^2(1) = 3.84$, $p = .05$, respectively. Specifically, co-located teams tended to rely on the unanimous strategy less than what would be expected by chance and relied on the majority strategy more than what would be expected by chance when reaching consensus on the non-repeated situation awareness queries. On the other hand, distributed teams tended to use the unanimous strategy more and the majority strategy less than what would be expected by chance.

Multiple regression analyses assessed whether the strategies used by teams in reaching their holistic response to the non-repeated situation awareness queries were predictive of their performance. The pattern of strategies used failed to significantly predict performance, $F(3, 16) < 1$.

Next we conducted similar analyses on the repeated situation awareness queries of Experiments AF3 and AF4. Recall that all six categories in Table 131 were used to classify the responses to the more quantitative repeated queries. As Table 131 shows, many teams utilized multiple

strategies in reaching consensus to the repeated query, as indicated by multiple high proportions. Unlike with the analyses for the non-repeated queries presented above, the set of possible strategies used in reaching consensus on the repeated query are not mutually exclusive (i.e., proportions for all strategies do not sum to 1). For example, using the unanimous strategy may inherently include the average strategy if all individual responses were the same.

It appears that co-located and distributed teams used the strategies in a very similar manner, relying upon the average strategy for the most part, followed by the leader strategy. Overall, co-located teams relied on each role when using a single individual's response; however, when distributed teams relied on the strategy to go with the single individual's response, it was usually the DEMPC's or PLO's response that was given as the team response.

Table 131

*Mapping Individual Responses to Team Responses on the Repeated Situation Awareness Queries for Co-located and Distributed Teams of Experiment AF3**

Team	Unanimous	Majority	Middle	Average	None	Leader	A	P	D	N
Co-located Teams										
1	0.43	0.14	0.29		0.00	0.14	0.00	0.00	0.14	7
2	0.29	0.29	0.14		0.00	0.29	0.00	0.29	0.00	7
3	0.29	0.43	0.00		0.00	0.29	0.00	0.00	0.29	7
7	0.14	0.14	0.14	0.29	0.00	0.57	0.00	0.29	0.29	7
8	0.00		0.17	0.17	0.00	0.33	0.17	0.17	0.00	6
11	0.14	0.29	0.14		0.00	0.43	0.00	0.29	0.14	7
12	0.14	0.14	0.00		0.00	0.71	0.29	0.14	0.29	7
13	0.14	0.29	0.00	0.43	0.00	0.57	0.29	0.14	0.14	7
14	0.14	0.29	0.00	0.29	0.00	0.57	0.00	0.14	0.43	7
16	0.29	0.43	0.00		0.00	0.29	0.00	0.14	0.14	7
Distributed Teams										
4	0.14	0.29	0.00	0.43	0.00	0.57	0.14	0.14	0.29	7
5	0.14	0.14	0.14		0.00	0.57	0.29	0.14	0.14	7
6	0.14	0.43	0.00		0.00	0.43	0.14	0.14	0.14	7
9	0.00	0.29	0.43		0.00	0.29	0.00	0.00	0.29	7
10	0.29		0.00		0.00	0.29	0.00	0.14	0.14	7
15	0.17	0.17		0.33	0.00	0.00	0.00	0.00	0.00	6
17	0.29	0.14	0.29		0.00	0.29	0.00	0.14	0.14	7
19	0.00	0.14	0.14	0.57	0.00	0.71	0.14	0.14	0.43	7
20	0.14	0.57	0.00		0.00	0.29	0.14	0.14	0.00	7
21	0.00	0.00	0.00	0.00	0.00	1.00	0.14	0.86	0.00	7

*Highlighted cells indicate which strategies teams used most.
A = AVO P = PLO D = DEMPC

Further analyses were not conducted to examine whether co-located and distributed teams relied on different strategies to reach consensus on the repeated situation awareness query because, as stated above, the strategies are defined in such a way that it is not meaningful to think of them as exclusive from one another. However, multiple regression analyses were used to assess whether

the strategies used by teams were predictive of their performance. The pattern of strategies used failed to predict performance, $F(5, 14) < 1$.

Finally, the analysis of repeated situation awareness queries was repeated for Experiment AF4. As Table 132 shows, co-located and distributed teams used the strategies in a very similar manner, relying upon the average strategy for the most part, followed by the leader strategy. Overall, co-located teams relied on each role when using the leader strategy; however, when distributed teams relied on the leader strategy, it was usually the DEMPC's or AVO's response that was given as the team response.

Table 132
*Mapping Individual Responses to Team Responses on the Repeated Situation Awareness Queries for Co-located and Distributed Teams of Experiment AF4**

Team	Unanimous	Majority	Middle	Average	None	Leader	A	P	D	N
Co-located Teams										
4	0.20	0.20	0.00	0.40	0.00		0.20	0.20	0.20	5
5	0.60	0.00	0.00		0.00	0.40	0.20	0.20	0.00	5
6	0.40	0.40	0.00		0.00	0.20	0.20	0.00	0.00	5
8	0.20	0.00			0.00		0.20	0.20	0.00	5
9	0.00		0.20		0.00		0.20	0.00	0.20	5
10	0.40	0.20	0.00		0.00	0.40	0.20	0.20	0.00	5
12	0.40	0.20	0.20		0.00	0.20	0.20	0.00	0.00	5
15	0.20	0.00	0.20	0.40	0.00		0.00	0.40	0.20	5
16	0.20	0.40	0.20		0.00	0.20	0.20	0.00	0.00	5
18	0.60	0.20	0.00		0.00	0.20	0.20	0.00	0.00	5
Distributed Teams										
1	0.20	0.20	0.00		0.00		0.20	0.20	0.20	5
2	0.20	0.20	0.00		0.00		0.40	0.00	0.20	5
3	0.20	0.40	0.20		0.00	0.20	0.00	0.20	0.00	5
7	0.00	0.40	0.00	0.20	0.20	0.40	0.40	0.00	0.00	5
11	0.40	0.20	0.20		0.00	0.20	0.00	0.00	0.20	5
13	0.40	0.40	0.20		0.00	0.00	0.00	0.00	0.00	5
14		0.20	0.20		0.00	0.20	0.20	0.00	0.00	5
17	0.20	0.20	0.20		0.20	0.20	0.00	0.00	0.20	5
19	0.00	0.00	0.20	0.20	0.00		0.40	0.00	0.40	5
20	0.00	0.40	0.00	0.20	0.00		0.20	0.00	0.40	5

*Highlighted cells indicate which strategies teams used most.
A = AVO P = PLO D = DEMPC

Again, further analyses were not conducted to examine whether co-located and distributed teams relied on different strategies to reach consensus on the repeated SA query due to the potential overlap in strategies used.

Multiple regression analyses were used to assess whether the strategies teams used were predictive of their performance. The pattern of strategies used significantly predicted performance, $F(5, 14) = 3.94$, $p < .02$. Semi-partial correlations revealed that the use of the

unanimous and leader strategies in reaching consensus on the repeated query was associated with lower performance, sr (14) = -.41, $p < .05$ and sr (14) = -.32, $p < .10$, respectively.

Overall, teams used the averaging strategy when responding to the repeated query while the unanimous and majority strategies were most utilized in reaching consensus on the non-repeated strategies. For both types of queries, teams appeared to use leader strategy a considerable proportion of time. However, for the non-repeated queries, in which the correct answer was known at the time the query was administered, teams clearly relied on the DEMPC's response the majority of the time. In contrast, when the leader strategy was chosen to reach consensus on the repeated query, no team member stood out as being the one the team relied. Perhaps teams felt that all team members had an equal say in the response to the repeated query, for which the correct answer was unknown at the time the query was administered, requiring the team to make a prediction. This finding also speaks to the success of the repeated query in terms of prediction team performance and sensitivity to the experimental manipulations. Perhaps the repeated query is a better indicator of team situation awareness (or some construct related to team performance) because it does not rely on the input of a single team member. Also interesting is the fact that strategy use was predictive of performance in Experiment AF4. Better teams used averaging, rather than the unanimous or leader strategies.

Taskwork knowledge. In order to identify strategies that the teams used to come to consensus in this rating task, the three individual and one team rating for each of the 55 concept pairs were examined for each of the twenty teams at Knowledge Sessions 2 for AF3 and for the session in AF4. For each pair, the set of four ratings was classified according to one of the first five rules listed in Table 133.

Results for AF3 Knowledge Session 2 are presented in Table 133. Some teams had missing or incomplete data and were excluded from the calculations. The table illustrates that co-located teams primarily used the majority rules and leader strategies. When using the leader strategy, co-located teams tended to chose the PLO's response as the team response. Distributed teams primarily used the majority rules strategy in reaching consensus on the taskwork measure. They also relied upon the leader strategy to a lesser degree and in those cases chose the PLO's response as the team response.

A two-way chi square analysis of teams at AF3, Knowledge Session 2 revealed a significant interaction between the proportion of times each strategy was used and whether teams were in the co-located or distributed condition, χ^2 (4) = 38.35, $p<.10$. One-way *post hoc* chi square tests showed that the unanimous strategy, χ^2 (1) = 18.70, $p<.10$, the leader strategy, χ^2 (4) = 18.70, $p<.10$, and the majority strategy, χ^2 (1) = 7.13, $p<.10$, were used differently by the two conditions. Co-located teams used the unanimous strategy more than expected while distributed teams used the unanimous strategy less than expected. Co-located teams also relied on one team member's answers more than expected while the distributed teams again used it less than expected. However, distributed teams used the majority strategy more than expected and the co-located teams used it less than expected.

Multiple regression analyses assessed whether any of the strategies used by co-located or distributed teams during the taskwork measure were predictive of performance (across all

missions, low workload missions, and high workload missions). For co-located teams, the proportion of times each strategy was used failed to predict performance for all missions, $F(4, 4) = 1.30$, low workload missions, $F(4, 4) = 1.62$, and high workload missions, $F(4, 4) = 1.91$. The same holds true for the distributed teams for all missions $F(4, 4) < 1$, low workload missions, $F(4, 4) < 1$, and high workload missions, $F(4, 4) = 1.05$.

Table 133

Mapping Individual to Team Responses for AF3, Knowledge Session 2 Taskwork Ratings *

Team	Unanimous	Majority	Middle	None	Leader	AVO	PLO	DEMPC
Co-located Teams								
1	.14	.34	.07	.00		.13	.22	.09
2	.31		.02	.00	.29	.11	.18	.00
3	.09	.22	.04	.20		.07	.29	.09
8	.29		.00	.00	.24	.02	.00	.22
11	.14		.05	.00	.38	.05	.24	.09
12	.24	.29	.07	.07		.04	.14	.14
13	.14		.04	.11	.34	.09	.22	.04
14		.34	.00	.00	.24	.13	.02	.09
16	.24	.27	.07	.05		.13	.14	.09
Distributed Teams								
4	.11		.02	.00	.31	.04	.22	.05
5	.09		.07	.00	.31	.05	.20	.05
6	.09		.04	.00	.42	.04	.24	.14
9	.16		.04	.02	.33	.11	.13	.09
10	.14		.00	.00	.40	.07	.29	.04
15	.22	.38	.00	.00		.11	.25	.04
19	.14		.02	.00	.34	.22	.09	.04
20	.22	.29	.04	.02		.14	.25	.04
21	.09		.00	.00		.13	.25	.07

*Highlighted cells indicate which strategies teams used most.

This analysis was repeated for the taskwork rating responses in Experiment AF4, however, the category, "middle" was dropped from this analysis. Results for co-located and distributed teams are presented in Table 134. As the table illustrates, the co-located teams used the majority rules or leader strategies. When teams relied on the leader strategy, it was usually the AVO's response that was given as the team response. The majority of distributed teams used the majority rules strategy with two teams using the leader strategy. In those two cases, it was either the PLO's or the DEMPC's response that was chosen as the team response.

Table 134

*Mapping Individual Responses to Team Responses for AF4 Taskwork Ratings**

Team	Unanimous	Majority	None	Leader	AVO	PLO	DEMPC
Co-located Teams							
4	.07	.33	.24		.11	.22	.04
5	.05		.18	.35	.07	.18	.09
6	.35		.02	.20	.05	.00	.15
8	.13	.38	.07		.16	.15	.11
9	.24		.04	.24	.09	.07	.07
10	.25		.18	.25	.04	.13	.09
12	.20		.11		.22	.04	.09
15	.04	.36	.16		.11	.25	.07
16	.15		.13	.35	.05	.20	.09
18	.22	.33	.11		.16	.05	.13
Distributed Teams							
1	.07		.11	.36	.11	.16	.09
2	.24		.04	.35	.13	.18	.04
3	.20		.02	.38	.15	.18	.05
7	.24		.07	.25	.11	.11	.04
11	.05	.31	.07		.11	.33	.13
13	.13	.27	.20		.07	.07	.25
14	.15		.07	.29	.11	.13	.05
17	.11		.15	.33	.07	.09	.16
19	.04		.07	.44	.05	.07	.31
20	.04		.22	.31	.07	.09	.15

*Highlighted cells indicate which strategies teams used most.

Overall, a two-way chi square analysis indicated that the proportion of times each strategy was used depended on whether teams were co-located or distributed, $\chi^2(3) = 6.40$, $p < .10$. One-way *post hoc* chi square tests indicated that it was the unanimous strategy that was used differently by co-located and distributed teams, $\chi^2(1) = 3.56$, $p < .10$, where co-located teams used the strategy more than what was expected by chance and distributed teams used the unanimous strategy less than expected.

Multiple regression analyses assessed whether any of the strategies used by co-located or distributed teams during the taskwork measure were predictive of performance (across all missions, low workload missions, and high workload missions). For co-located teams, the proportion of times each strategy was used was able to marginally predict performance for all missions, $F(4, 5) = 3.68$, and high workload missions, $F(4, 5) = 3.79$. The proportions were not able to predict performance for low workload missions, $F(4, 5) = 2.55$. For distributed teams, The proportion of times each strategy was used was unable to predict performance for all missions $F(4, 5) = 1.12$, low workload missions, $F(4, 5) = .951$, and high workload missions, $F(4, 5) = 1.73$.

Teamwork knowledge. In order to identify strategies that the teams used to come to consensus during the teamwork-measure, the three individual responses and one team response to each of

the 16 questions in the measure were examined for each of the twenty teams. For each response, the set of four responses (i.e., 3 individual and 1 team response) was classified according to one of the first four rules in Table135. Results for co-located and distributed teams at AF3, Knowledge Session 2 are presented in Table 137. The table illustrates that all the teams used a unanimous decision making strategy for a greater proportion of time than the other three strategies.

Table 135
*Classification of AF3, Knowledge Session 2 Teamwork Responses on the Basis of Mapping Individual to Team Responses**

Team	Unanimous	Majority	None	Leader	AVO	PLO	DEMPC
Co-located Teams							
1		.06	.00	.38	.06	.19	.12
2		.12	.00	.06	.06	.00	.00
3		.31	.00	.06	.00	.00	.06
7		.31	.00	.06	.00	.00	.06
8		.06	.00	.06	.00	.00	.06
11		.06	.00	.06	.00	.06	.00
12		.44	.00	.00	.00	.00	.00
13		.19	.00	.00	.00	.00	.00
14		.19	.12	.12	.12	.00	.00
16		.19	.00	.00	.00	.00	.00
Distributed Teams							
4		.06	.00	.12	.06	.19	.12
5		.19	.00	.06	.06	.00	.00
6		.12	.00	.00	.00	.00	.00
9		.12	.00	.19	.00	.06	.12
10		.25	.00	.00	.00	.00	.00
15		.31	.00	.06	.06	.00	.00
17		.31	.00	.00	.00	.00	.00
19		.00	.06	.06	.00	.00	.06
20		.37	.06	.25	.19	.00	.06
21		.06	.00	.00	.00	.00	.00

*Highlighted cells indicate which strategies teams used most

A two-way chi square analysis indicated that the proportion of times each strategy was used at Knowledge Session 2 did not depend on whether the teams were in the co-located or distributed condition, χ^2 (3) = .15.

A multiple regression analysis was conducted to determine whether the pattern of strategies used was predictive of performance. The proportion of times co-located teams used each strategy failed to predict performance across all missions, $F(3,6) = 2.14$, low workload missions, $F(3, 6) = 1.03$, and high workload missions, $F(3, 6) = 1.38$. The proportion of times distributed teams used each strategy also failed to predict performance across all missions, $F(3, 6) < 1$, low workload missions, $F(3, 6) < 1$, and high workload missions, $F(3, 6) < 1$.

The analysis was repeated for Experiment AF4 data. Results for co-located and distributed teams are presented in Table 136. As the table illustrates, it appears that most co-located teams used the unanimous strategy or the majority rules strategy. One team used the leader strategy and they based their team answers on the PLO's responses. Distributed teams tended to rely on the unanimous strategy and, to a lesser degree, the majority rules strategy.

Table 136
*Classification of AF4 Teamwork Responses on the Basis of Mapping Individual to Team Responses**

Team	Unanimous	Majority	Leader	None	AVO	PLO	DEMPC
Co-located Teams							
4	.31	.25		.00	.00	.44	.00
5			.13	.00	.00	.06	.06
6		.31	.31	.00	.31	.00	.00
8	.44		.00	.00	.00	.00	.00
9	.38		.13	.00	.13	.00	.00
10		.38	.06	.00	.00	.00	.06
12	.25		.13	.13	.13	.00	.00
15		.38	.13	.00	.13	.00	.00
16		.25	.31	.00	.06	.19	.06
18		.13	.25	.00	.19	.00	.06
Distributed Teams							
1	.38		.00	.00	.00	.00	.00
2		.19	.06	.00	.00	.06	.00
3		.38	.06	.06	.00	.06	.00
7		.31	.06	.00	.00	.06	.00
11		.31	.13	.00	.00	.06	.00
13		.25	.19	.00	.06	.06	.06
14			.00	.00	.00	.00	.00
17		.25	.13	.00	.00	.13	.00
19	.19		.19	.06	.06	.00	.13
20		.13	.06	.00	.00	.00	.06

*Highlighted cells indicate which strategies teams used most.

A two-way chi square analysis revealed a significant interaction between the proportion of times each strategy was used and whether teams were in the co-located or distributed condition, χ^2 (4) = 8.26, $p<.10$. One-way *post hoc* chi square tests showed that the proportion of time teams used the leader strategy differed between the two conditions, χ^2 (1) = 5.90, $p<.10$. Co-located teams used this strategy more than expected while distributed teams used the strategy less than expected.

Multiple regression analyses assessed whether any of the strategies used by co-located or distributed teams during the teamwork measure were predictive of performance (across all missions, low workload missions, or high workload missions). The proportion of times each strategy was used by co-located failed to predict performance for all missions, $F(3, 6) = 1.58$, low workload missions, $F(3, 6) = 1.53$, and high workload missions, $F(3, 6) = 1.71$. The same

holds true for the distributed teams for all missions, $F(3, 6) < 1$, low workload missions, $F(3, 6) < 1$, and high workload missions, $F(3, 6) < 1$.

Summary. These results are exploratory and demonstrate that team process can be extracted from the pattern of responses in individual and consensus-based knowledge tasks. The function required to map from individual to team responses provides an estimate of team process behavior. Some differences emerged in the predominant strategy used across the different knowledge tasks (situation awareness, taskwork, teamwork) and different conditions. For repeated situation awareness queries in which the responses are quantitative in nature teams tended to use an averaging strategy. In other cases teams seemed to prefer a unanimous or majority rules strategy. In some cases, however, teams who performed better were less likely to use the majority rule strategy and more likely to rely on a leader. This was especially true for distributed teams.

4.20 Archival Analysis to Evaluate Measures: Discussion

The analyses documented in this section were aimed at evaluating the measures used in this project in terms of reliability and validity. Results indicated adequate reliability for all of our primary measures with the exception of teamwork knowledge. Data available to assess teamwork knowledge reliability was sparse, but nonetheless provided weak support. In terms of validity, we tested the relative ability of our knowledge measures and critical incident process to predict team performance. Our situation awareness measure was the best predictor followed by our taskwork knowledge measure. However, In Experiments AF3 and AF4 of this project we have generally found that our knowledge measures are only weakly related to performance. Further, the MTMM analysis indicated that the taskwork knowledge and teamwork knowledge measures have low construct validity. We believe that previous strong correlations obtained in AF1 were due to the placement of the knowledge elicitation session. Placing it prior to mission experience is probably too soon for reliable knowledge assessment, and placing it at the end of the experiment is probably too late with regard to fatigue. It may be that early on, but not too early in the experiment, the amount of taskwork knowledge is indicative of team performance. After this point other factors become more important. More generally, we are accumulating evidence that suggests that variations in teamwork and taskwork knowledge beyond the criterion level set at training are not strongly tied to later variations in team performance.

We also conclude that our situation awareness measure (repeated query) works very well and is predictive of performance, but perhaps for the wrong reasons. That is, although this measure has predictive validity we are suspicious of its construct validity. To do well on the repeated query one simply needs to know how many targets there are in a mission and whether his or her team is capable of photographing them all. After a few low workload missions, this should become clear.

Finally, we took a deep look at collective versus holistic measures. In our first analysis we showed that the rank ordering of teams on the basis of performance depends on whether that performance is assessed collectively (averaging individual scores) or holistically (a composite team performance score as we have). We further showed that including holistic measures in our regression analysis adds to the models ability to predict team performance. This analysis

suggested that both forms of measure are important. In our final analysis we demonstrated how team process behaviors in the form of consensus decision strategies can be inferred from patterns of responding to individual and group-level knowledge probes. In some cases the process behaviors extracted differed for co-located versus distributed teams. In other cases the frequencies of particular categories of process behavior were predictive of team performance.

4.21 Conclusions

This report summarizes a three-year effort that included three experiments and two additional sets of data analyses, which encompassed data from two studies prior to this effort. Overall our results speak to three different topics touching respectively on applied, theoretical, and methodological issues central to team cognition: 1) distributed mission environments, 2) team cognition in command-and-control, and 3) measuring team cognition. We discuss each of these in turn.

4.21.1 Distributed Mission Environments

Results from two studies in which teams were either co-located (in the same room talking over headsets) or distributed (in different rooms talking over headsets) indicated no performance effects of geographic dispersion. This result needs to be qualified with the fact that the study was conducted in the UAV-STE and in both cases communication occurred over headsets. Thus this test should not be confused with a test of face-to-face versus distributed team interaction; rather, to a large extent, all information coordination was moderated in a like manner. On the other hand, the UAV-STE is based on actual USAF Predator operations and our manipulation is relevant to this setting in which operators sit side-by-side and converse over head-sets. Further we suspect that the task in our UAV-STE is representative of many command-and-control tasks and we speculate that our findings would generalize to these settings. For instance, holding mode of communication constant, subtle differences in team processes may obtain by introducing geographic dispersion. To the extent that this did not impact team performance in our UAV-STE indicates that such differences remain subtle. However this is not to say that patterns will not accrue over time, eventually impacting team performance. This is an important issue to consider in designing a command-and-control task, in that many real-world tasks have a life span greater than seven 40-minute missions. Ultimately however, specific tasks will obviate their own considerations over time, and specifically we would not recommend drawing conclusions about other very different tasks such as collaborative design or process control.

In terms of co-located/distributed performance we must also be cautious when drawing conclusions on the basis of null results. Lack of effect is often due to inadequate statistical power, especially when research is conducted on specialized populations (i.e., teams of 3) where sampling requirements are constraining. Indeed our studies have a very modest sample size, however, it was adequate to reveal other effects such as the effect of increased workload; a robust effect that was common across both experiments. In addition, we note that in both experiments there is, in fact, a tendency for distributed teams to outperform co-located teams suggesting that if there is a difference it is in the opposite direction than expected. This raises the prospect that it is in fact the DME teams who have the advantage and co-located teams who play catch-up, especially under high workload. Finally, we have also carefully explored the data

for outliers and for effects on specific components of the performance score. In any case, we are confident in our conclusion that the DME was not detrimental to team performance.

Although the DME was not detrimental to team performance it did have some subtle effects on knowledge and process. We say "subtle" because evidently the taskwork knowledge deficits and process deficits experienced by distributed teams did not have an impact on performance. Alternatively, another possibility that we currently entertain is that teams adapt to their specific situation through varying team process behaviors. Teams in DMEs may experience a loss of taskwork knowledge and compensate for this loss through different patterns of team interaction. For instance, team members in a DME who lack interpositional knowledge, might compensate by relying more on the information of other team members, thus requiring a different pattern or form of team process. For example, by not taking advantage of post mission opportunities to discuss and evaluate performance, or explicitly noting emergent properties of the task, DME teams can overcome this by applying more effort during the missions in terms variable communication patterns (Kiekel, Gorman, & Cooke, 2004). This pattern of results is interesting and suggests that there may very well be a detrimental effect of DMEs in a task in which certain types of knowledge (i.e., taskwork knowledge) and certain types of process behaviors (i.e., communication over a fixed medium; Gorman, Cooke, & Kiekel, 2004) are critical for performance. On the other hand our analysis of workload suggests that DMEs may be less stressful to team members at least in terms of perceived workload. Specifically, distributed DEMPCs perceived less pressure imposed by teammates, less pressure to perform well, and less time pressure than co-located DEMPCs.

Overall, the lack of a performance effect is good news for the Air Force and other military agencies for which network centric warfare is taking a front seat to co-located collaboration. At least in a command-and-control setting, teams seem to be able to thrive in this environment. Additionally, a number of critical observations made in this report have pointed to the necessity of examining other factors, including knowledge and team process, when considering command-and-control task design and long term coordination effects.

4.21.2 Team Cognition in Command and Control

What have we learned about team cognition in a command-and-control setting? First, we see a repeated and robust pattern of team skill acquisition across all of the experiments that we have conducted in the UAV-STE. Teams need to interact for approximately four 40-minute missions before they reach asymptotic levels of performance. Second, teams have mastered their individual tasks (to criterion) in the individual training that preceded the first mission. Although certainly there is some further development of individual skills, we believe that in large part what they are developing are *team* skills. The question for the future is what exactly is it that teams are learning (as teams) during this period? We achieved a partial glimpse at the answer when we ran the expert teams of Experiment 3. Several of these teams had rapidly increased task acquisition relevant to an average UAV-STE team. The only difference between these two types of teams was familiarity working together (although the Experimenter expert team also had task familiarity). This is therefore apparently a very important issue to consider when constructing *ad hoc* command-and-control teams as is often the practice in military settings.

We have additionally found that the development of team process and team situation awareness parallels the performance curve. On the other hand, we have had only weak evidence that further development of taskwork and teamwork knowledge (beyond training) is important for team performance. When looking at individual characteristics that are predictive of team performance, taskwork and teamwork knowledge are not implicated. Instead, good teams have individuals with good working memory, high GPAs, and good situation awareness. So, the individuals on a good team are generally smart, but not particularly knowledgeable about the task in a declarative sense. Again, looking at our teams from the benchmarking experiment (Experiment 3), it seems that the best teams demonstrate good process behaviors, as opposed to superior knowledge. Putting all of these findings together, we believe that teams are learning how to interact or coordinate. Thus they must learn how to push and pull information in the appropriate direction at the appropriate time. They must develop good team process. Therefore in our scheme of team cognition where we once focused on team knowledge, we are now focused on team process. Furthermore, we see team process as cognitive processing that occurs at the team level. Unlike individual tasks, it is this cognitive processing that largely has to be developed in teams; the knowledge already exists in the world of the task, where that knowledge comes from and where it needs to go are the defining processes that good teams develop by interacting over time.

4.21.3 Measuring Team Cognition

Much of our effort has been devoted to developing and evaluating measures of team cognition. Based on our preceding analysis and discussion, we will be spending more time in the future developing and evaluating metrics of process or coordination. In this effort, however, we focused on knowledge measures. We have already discussed some issues with our taskwork and teamwork measures. Teamwork was not adequately reliable and both measures had problems with validity. Placement of the elicitation session was also an issue. It is also possible that our measures are adequate but that there is little relation between the knowledge construct and team performance in reality. That is, as discussed previously, the knowledge is already there, in order to become functional knowledge however, it must be tapped through team coordination and other good process behaviors. The quadratic relationship between teamwork knowledge and team performance found in Experiment 2 further supports the idea that after some prerequisite level of teamwork knowledge has been attained, further knowledge development is negatively correlated with team performance. Teams should instead be focused on the development of coordination skill.

Our situation awareness measure on the other hand, did much better, though perhaps for the wrong reasons. One difference between the situation awareness measure and the knowledge measure was that the former was taken *during* mission performance and in fact embedded within the task. This may increase sensitivity to finding differences in mission performance in that it occurs within the flow of team coordination, rather than static knowledge of elements in the task. Thus one important challenge is to identify on-line and task embedded means to elicit team knowledge during task performance. We have also been involved in some communication analysis work that is promising in this regard, but may be more in line with team process elicitation than knowledge.

Our major measurement contribution in this effort was in the development and evaluation of holistic measures of knowledge. We were able to show that holistic and collective measures result in different outcomes and each provide metrics that account for variance in team performance. For example, in the Experiment 2 knowledge session, we found that distributed teams had equivalent taskwork knowledge to co-located teams only when measured holistically (i.e., they had poor taskwork knowledge unless they were allowed to come to consensus on a rating). In other words distributed teams had a hard time tapping this knowledge without interaction. Thus, in this case only a holistic metric would have been able to adequately characterize the distributed true, effective knowledge of the task. Again these observations anticipate a need for on-line holistic assessment in order to truly tap the effective knowledge of teams. Our work in communication analysis is in line with this need.

Finally, we also have demonstrated the importance of considering the role of team members when it comes to measuring team performance or cognition on a heterogeneous team. Whether the individual was an AVO, PLO, or DEMPC mattered when it came to relationships between working memory capacity or grade point average and team performance. The analysis of workload effects (see Appendix **) also points to the impact of team role. For instance, DEMPC performance, like team performance, was consistently hindered by the high workload scenario, while AVO performance was not affected by workload. Most of these methodological conclusions are based on the premise that a team, especially one that is heterogeneous, is more than the sum of the individual team members.

4.21.4 Summary

Overall the efforts reported in this tech report can be summarized by three major conclusions. First, we have addressed the applied issue concerning potential ill effects of distributed mission environments central to network centric warfare. We found no performance differences between co-located and distributed teams. However subtle differences in terms of both team process and individual taskwork knowledge were found, suggesting that distributed equivalence in performance may be achieved via modified team process strategies and ultimately differential mechanisms for transforming individual knowledge into effective knowledge. These results begin to suggest the importance of considering long-term process behaviors that can accrue over time ultimately impacting team performance in command-and-control task and training design.

Second, we have made methodological progress and confirmed suspicions that on-line embedded and holistic measures are often the most appropriate when assessing team knowledge. The very nature of team command-and-control tasks necessitates the inclusion of ongoing, interaction-based elicitation methods when evaluating team member knowledge. That is to say, team cognition is dynamic and difficult if not impossible to locate in the static knowledge within an individual's head. This individual, static knowledge is devoid of what we believe to be at the heart of team knowledge-team member interaction; it is this *team* knowledge that we believe is best captured by on-line, holistic elicitation methods. For example, we believe such methods are required in order to effectively establish knowledge transfer as it occurs in practice when evaluating new team task environments or team training regimes.

Third, and in a similar vein, our theoretical views of team cognition have evolved. We feel that we have established that the acquisition of coordinative team skills lies at the heart of what it means to become proficient as a team. As a research team we have accordingly shifted our focus to the study of the acquisition as well as retention of these coordinative skills. We believe that ultimately models useful for team training and team command-and-control design must focus on the development and retention of these coordination skills as coordination lies at the heart of team cognition as well as differences between co-located and distributed command-and-control UAV teams.

4.22 References

Barnes, M. J., Knapp, B. G., Tillman, B. W., Walters, B. A., & Velicki, D. (2000). *Crew Systems Analysis of Unmanned Aerial Vehicle (UAV) Future Job Tasking Environments.* (Army Research Laboratory Technical Report, ARL-TR-2081).

Braun, C. C., Bowers, C. A., Holmes, B. E., & Salas, E. (1993). Impact of task difficulty on the acquisition of aircrew coordination skills. *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting,* 1262-1266.

Brehmer, B. & Dörner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Computers in Human Behavior, 9,* 171-184.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix, *Psychological Bulletin, 56,* 81-105.

Cannon-Bowers, J. A., Burns, J. J., Salas, E., & Pruitt, J. S. (1998). Advanced technology in scenario-based training. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making Decisions Under Stress,* (pp. 365-374). Washington, DC: American Psychological Association.

Cannon-Bowers, J. A., & Salas, E. (1997). Teamwork competencies: The interaction of team member knowledge skills and attitudes. In O. F. O'Neil (Ed.), *Workforce readiness: Competencies and assessment* (pp. 151-174). Hillsdale, NJ: Erlbaum.

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In J. Castellan Jr. (Ed.), Current *issues in individual and group decision making* (pp. 221-246). Hillsdale, NJ: Erlbaum.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Cognitive skills and their acquisition* (pp. 141-189). Hillsdale, NJ: Erlbaum.

Contractor, N. S., Seibold, D. R., & Heller, M. A. (1996). Interactional influence in the structuring of media use in groups: Influence of members' perceptions of group decision support system use. *Human Communication Research, 22,* 451-481.

Cooke, N. J., Kiekel, P. A., Bell, B., & Salas, E. (2002). Addressing limitations of the measurement of team cognition. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting,* 403-407.

Cooke, N. J., Kiekel, P. A., & Helm E. (2001). Measuring team knowledge during skill acquisition of a complex task. *International Journal of Cognitive Ergonomics: Special Section on Knowledge Acquisition, 5,* 297-315.

Cooke, N. J., Rivera, K., Shope, S. M., & Caukwell, S. (1999). A synthetic task environment for team cognition research. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting,* 303-307.

Cooke, N. J., Salas, E., Cannon-Bowers, J. A., & Stout, R. (2000). Measuring team knowledge. *Human Factors, 42,* 151-173.

Cooke, N. J. & Shope, S. M. (1998). *Facility for cognitive engineering research on team tasks.* (Report for Grant No. F49620-97-1-0149, submitted to AFOSR, Bolling AFB, Washington, DC).

Cooke, N. J., & Shope, S. M. (2002a). The CERTT-UAV task: A synthetic task environment to facilitate team research. *Proceedings of the Advanced Simulation Technologies Conference: Military, Government, and Aerospace Simulation Symposium,* (pp. 25-30). San Diego, CA: The Society for Modeling and Simulation International.

Cooke, N. J., & Shope, S. M. (2002b). Behind the scenes. *UAV Magazine, 7,* 6-8.

Cooke, N. J., & Shope, S. M. (in press). Designing a synthetic task environment. In S. G. Schiflett, L. R. Elliott, E. Salas, &. Coovert, M. D., *Scaled Worlds: Development, Validation, and Applications*. Surrey, England: Ashgate.

Cooke, N. J., Shope, S. M., & Kiekel, P. A. (2001). *Shared-knowledge and team performance: A cognitive engineering approach to measurement.* (Technical Report for AFOSR Grant No. F49620-98-1-0287).

Cooke, N. J., Shope, S.M., & Rivera, K. (2000). Control of an uninhabited air vehicle: A synthetic task environment for teams. *Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting*, 389.

Cooke, N. J., Stout, R., Rivera, K., & Salas, E. (1998). Exploring measures of team knowledge. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, 215-219.

Cooke, N. J., Stout, R., & Salas, E. (1997) Expanding the measurement of situation awareness through cognitive engineering methods. *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting,* 215-219.

Cooke, N. J., Stout, R.J., & Salas, E. (2001). A knowledge elicitation approach to the measurement of team situation awareness. In M. McNeese, M. Endsley, & E. Salas, (Eds.), *New Trends in Cooperative Activities: System Dynamics in Complex Settings,* (pp. 114-139). Santa Monica, CA: Human Factors.

Crowne, D. & Marlowe, D. (1964). *The approval motive: studies in evaluative dependence.* Wiley, New York.

Dolgin, D., Kay, G., Wasel, B, Langelier, M, & Hoffmann, C. (2001). Identification of the cognitive, psychomotor, and psychosocial skill demands of Uninhabited Combat Air Vehicle (UCAV) operators. *Journal of Survival and Flight Equipment, 30*, 219-225.

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J. M., Nikolic, D., & Manning, C. A. (1998). Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly, 6,* 1-20.

Fatolitis, P. (2003). Pioneer UAV selection battery validation. Unpublished Manuscript

Foushee, H. C. (1984). Dyads and triads at 35,000 feet. *American Psychologist, 39,* 885-893.

Fowlkes, J., Dwyer, D. J., Oser, R. L., & Salas, E. (1998). Event-based approach to training (EBAT). *The International Journal of Aviation Psychology, 8*, 209-221.

Glaser, R. & Chi, M. T. H. (1988). Overview. In M.T.H. Chi, R. Glaser, and M.J. Farr (Eds.), The *Nature of Expertise* (xv-xxviii). Hillsdale, NJ: Erlbaum.

Gorman, J.C., Cooke, N.J., & Kiekel, P.A. (accepted). Dynamical perspectives on team cognition. Proceedings paper accepted for *Human Factors and Ergonomics Society 48th annual meeting*.

Gugerty, L., DeBoom, D., Walker, R., & Burns, J. (1999). Developing a simulated uninhabited aerial vehicle (UAV) task based on cognitive task analysis: Task analysis results and preliminary simulator data. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 86-90). Santa Monica, CA: Human Factors and Ergonomics Society.

Gugerty, L., Hall, E., & Tirre, W. (1998). Designing a simulation environment for uninhabited aerial vehicle (UAV) operations based on cognitive task analysis. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (p. 1609). Santa Monica, CA: Human Factors and Ergonomics Society.

Hart, S.G. and Staveland, L.E. (1988). Development of the NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds.), *Human Mental Workload* (pp. 139-177). North-Holland: Elsevier Science Publishers.

Hedlund, J., Ilgen, D. R., & Hollenbeck, J. R. (1998). Decision accuracy in computer-mediated versus face-to-face decision-making teams. *Organizational and Behavior and Human Decision Processes, 76,* 30-47.

Heslin, R. (1964). Predicting group task effectiveness from member characteristics. *Psychological Bulletin, 62,* 248-256.

Hollenbeck, J. R., Ilgen, D. R., Sego, D. J., Hedlund, J., Major, D. A., & Phillips, J. (1995). Multilevel theory of team decision making: Decision performance in teams incorporating distributed expertise. *Journal of Applied Psychology, 80,* 292-316.

Hollenbeck, J. R., Moon, H, Ellis, A. P., West, B. J., Ilgen, D. R., Sheppard, L., Porter, C. O., & Wagner, J. A. (2002). Structural contingency theory and individual differences: Examination of external and internal person-team fit. *Journal of Applied Psychology, 87,* 599-606.

Howell, W. C., & Cooke, N. J. (1989). Training the human information processor: A look at cognitive models. In I. Goldstein (Ed.), *Training and Development in Work Organizations: Frontier Series of Industrial and Organizational Psychology, Volume 3,* (pp. 121-182). *New* York: Jossey Bass.

Hutchins, E. (1991). The social organization of distributed cognition. In L. B. Resnick, J. M. Levine, and S. D. Teasley (Eds.), *Socially Shared Cognition* (pp. 283-301). Washington, D.C.: American Psychological Association.

Kiekel, P.A., Gorman, J.C., & Cooke, N.J. (accepted). Measuring speech flow of co-located and distributed command and control teams during a communications channel glitch. Proceedings paper accepted for *Human Factors and Ergonomics Society 48th annual meeting*.

Kleinman, D. L., & Serfaty, D. (1989). Team performance assessment in distributed decision making. In R. Gilson, J. P. Kincaid, & B. Godiez (Eds.), *Proceedings: Interactive Networked Simulation for Training Conference* (pp. 22-27). Orlando, FL: Institute for Simulation and Training.

Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management, 20,* 403-437.

Kyllonen, P. C. (1995). CAM: A theoretical framework for cognitive abilities measurement. In D. Detterman (Ed.), *Current topics in human intelligence: Volume IV. Theories of intelligence* (pp. 307-359). Norwood, NJ: Ablex

Kyllonen, P. C. (1996). Is Working Memory Capacity Spearman's *g*? In I. Dennis & P. Tapsfield (Eds.) *Human Abilities: Their Nature and Measurement*. Mahwah, N.J.: Erlbaum.

Kyllonen, P.C., & Christal, R.E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14,* 389–433.

Leedom, D. K., & Simon, R. (1995). Improving team coordination: A case for behavior-based training. *Military Psychology, 7,* 109-122.

LePine, J.A., Hollenbeck, J.R., Ilgen, D.R., & Hedlund, J. (1997). Effects of individual differences on the performance of hierarchical decision-making teams: Much more than *g*. *Journal of Applied Psychology, 82,* 803-811.

Mantovani, G. (1996). *New communication environments from everyday to virtual*. Bristol, PA: Taylor & Francis.

Martin, E., Lyon, D. R., & Schreiber, B. T. (1998). Designing synthetic tasks for human factors research: An application to uninhabited air vehicles. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 123-127). Santa Monica, CA: Human Factors and Ergonomics Society.

Miller, C. S., Lehman, J. F., & Koedinger, K. R. (1999). Goals and learning in microworlds. *Cognitive Science, 23*, 305-336.

Orasanu, J. (1990). Shared mental models and crew decision making. (Tech. Rep. No. 46). Princeton, NJ: Princeton University, Cognitive Science Laboratory.

Postmes, T., & Spears, R. (1998). Deindividuation and anti-normative behavior: A meta-analysis. *Psychological Bulletin, 123*, 238-259.

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-effects of computer-mediated communication. *Communication Research, 25,* 689-715.

Prince, C., Chidester, T. R., Cannon-Bowers, J., & Bowers, C. (1992). Aircrew coordination – Achieving teamwork in the cockpit. In R. Swezey & E. Salas (Eds.), *Teams: Their training and performance* (pp. 329-353). New York: Ablex.

Prince, C. & Salas, E. (1993). Training and research for teamwork in the military aircrew. In E. L. Wiener, B. G. Kanki, and R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 337-366). San Diego, CA: Academic Press.

Robertson, M. M., & Endsley, M. R. (1997). Development of a situation awareness training program for aviation maintenance. *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting* (pp.1163-1167). Santa Monica, CA: Human Factors and Ergonomics Society.

Rogers, Y., & Ellis, J. (1994). Distributed cognition: An alternative framework for analyzing and explaining collaborative working. *Journal of Information Technology, 9,* 119-128.

Salas, E., Bowers, C. A., & Rhodenizer, L. (1998). It is not how much you have but how you use it: Toward a rational use of simulation to support aviation training. *The International Journal of Aviation Psychology, 8*, 197-208.

Salas, E. Cannon-Bowers, J.A., Church-Payne, S., & Smith-Jentsch, K. A. (1998). Teams and teamwork in the military. In C. Cronin (Ed.), *Military Psychology: An Introduction* (pp. 71-87). Needham Heights, MA: Simon & Schuster.

Salas, E. Dickinson, T. L., Converse, S. A., & Tannenbaum, S. I. (1992). Toward an understanding of team performance and training. In R. W. Swezey & E. Salas (Eds.), *Teams: Their training and performance* (pp. 3-29). Norwood, NJ: Ablex.

Schreiber, B. T., Lyon, D. R., Martin, E. L., & Confer, H. A. (2002). Impact of prior experience on learning predator UAV operator skills. AFRL Technical Report No. AFRL-HE-AZ-TR-2002-0026.

Schvaneveldt, R. W. (1990). Pathfinder Associative Networks: Studies in Knowledge Organization. Norwood, NJ: Ablex.

Stout, R., Cannon-Bowers, J. A., & Salas, E. (1996). The role of shared mental models in developing team situation awareness: Implications for training. *Training Research Journal, 2,* 85-116.

Steiner, I. D. (1972). *Group process and productivity.* New York: Academic Press.

Stout, R. J., Salas, E., & Carson, R. (1994). Individual task proficiency and team process: What's important for team functioning? *Military Psychology, 6*, 177-192.

Taylor, R.M. (1990). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In Situational Awareness in Aerospace Operations (AGARD-CP-478) (pp.3/1-3/17). Neuilly Sur Seine, France: NATO – AGARD.

Wickens, C. D., & Holland, J. G. (2000). Engineering Psychology and Human Performance. Upper Saddle River, NL: Prentice Hall.

Wilson, J. R. (2000). Network-centric warfare marks the frontier of the 21st century battlefield. *Military and Aerospace Electronics*, January, 13-30.

Zalesny, M. D., Salas, E., & Prince, C. (1995). Conceptual and measurement issues in coordination: Implications for team behavior and performance. In G. R. Ferris (Eds.), *Research in Personnel and Human Resources Management (Vol. 1,* pp. 81-115*)*. Greenwich, CT: JAI Press.

4.23 Acknowledgements

This work was supported by the Air Force Research Laboratory under agreements F49620-01-1-0261 and F49620-03-1-0024. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

This effort benefited from the aid of several individuals. Steven Shope of US Positioning developed the CERTT remote workstation, performed numerous upgrades to the CERTT experimenter station, and generated a variety of data collection and post processing software. Pat Fitzgerald and Rebecca Keith were both instrumental in conducting numerous data analyses. Thanks to Paulette Dutcher for her assistance with technical editing. Thanks to Christy Caballero for her assistance in assembling this report. Thanks to Brian Bell for his work on the working memory and processing speed measures, as well as the preliminary work on voice analysis.

5.0 PUBLICATIONS ASSOCIATED WITH THIS EFFORT

Journal Articles

1) Cooke, N. J., Kiekel, P. A., & Helm E. (2001). Measuring team knowledge during skill acquisition of a complex task. *International Journal of Cognitive Ergonomics: Special Section on Knowledge Acquisition, 5,* 297-315.
2) Cooke, N. J., Kiekel, P.A., Salas, E., Stout, R.J., Bowers, C., Cannon-Bowers, J. (2003). Measuring Team Knowledge: A Window to the Cognitive Underpinnings of Team Performance. *Group Dynamics: Theory, Research and Practice, 7,* 179-199.

Books and Book chapters

3) Cooke, N. J., Stout, R.J., & Salas, E. (2001). A knowledge elicitation approach to the measurement of team situation awareness. In M. McNeese, M. Endsley, & E. Salas, (Eds.), *New Trends in Cooperative Activities: System Dynamics in Complex Settings,* pp. 114-139. Santa Monica, CA: Human Factors.
4) Cooke, N. J., Salas, E., Kiekel, P. A., & Bell, B. (2004). Advances in measuring team cognition. In E. Salas and S. M. Fiore (Eds.), *Team Cognition: Understanding the Factors that Drive Process and Performance, pp. 83-106,* Washington, DC: American Psychological Association.
5) Cooke, N. J., & Shope, S. M. (in press), Designing a synthetic task environment. In S. G. Schiflett, L. R. Elliott, E. Salas, & M. D. Coovert, *Scaled Worlds: Development, Validation, and Applications*. Surrey, England: Ashgate.
6) Cooke, N. J. (in press). Measuring Team Knowledge. Handbook on Human Factors and Ergonomics Methods. Taylor Francis.
7) Cooke, N. J., & Shope, S. M. (in press). Synthetic Task Environments for Teams: CERTT's UAV-STE Handbook on Human Factors and Ergonomics Methods. Taylor Francis.
8) Bell, B., Cooke, N. J., Gorman, J., Kiekel, P. K., DeJoode, J., Pedersen, H., & Keith, R. (submitted). Distributed Mission Environments: Effects of Geographic Dispersion on Team Cognition and Performance. S. Fiore and E. Salas (Eds.), Where is the Learning in Distance Learning? Towards a Science of Distributed Learning and Training.

Proceedings

9) Cooke, N. J., Kiekel, P. A., & Helm E. (2001). Comparing and validating measures of team knowledge. *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting.*
10) Cooke, N. J., & Shope, S. M. (2002). The CERTT-UAV Task: A Synthetic Task Environment to Facilitate Team Research. *Proceedings of the Advanced Simulation Technologies Conference: Military, Government, and Aerospace Simulation Symposium,* pp. 25-30. San Diego, CA: The Society for Modeling and Simulation International.
11) Cooke, N. J., Kiekel, P. A., Bell, B., & Salas, E. (2002). Addressing limitations of the measurement of team cognition. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting,* 403-407.
12) Cooke, N. J. & Shope, S. M. (2002). Behind the scenes. *UAV Magazine, 7,* 6-8.

13) Bell, B. G., & Cooke, N. J. (2003). Cognitive ability correlates of performance on a team task. *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting*, 1087-1091.

Presentations

14) Cooke, N. J., Kiekel, P. A., & Helm E. (2001). Comparing and validating measures of team knowledge. Paper presented at 45th annual meeting of the Human Factors and Ergonomics Society, October 8-12, Minneapolis, MN.
15) Cooke, N. J., & Bell, B. (2001) The CERTT Lab: Cognitive Engineering Research on Team Tasks. Poster presented at the first annual NMSU Research and Creative Activities Fair, September 27, Las Cruces, NM.
16) Cooke, N. J., & Shope, S. M. (2002). The CERTT-UAV Task: A Synthetic Task Environment to Facilitate Team Research. Paper presented at the Advanced Simulations Technologies Conference, April 14-18, San Diego, CA.
17) Cooke, N. J., Kiekel, P. A., & Bell, B., & Salas, E. (2002). Addressing limitations of the measurement of team cognition. Paper presented at 46th annual meeting of the Human Factors and Ergonomics Society, September 30-October 4, Baltimore, MD.
18) Bell, B. G., & Cooke, N. J. (2003). Cognitive ability correlates of performance on a team task. Poster presented at 47th annual meeting of the Human Factors and Ergonomics Society, October 13-17, Denver, CO.

Invited Talks and Workshop Presentations

19) Cooke, N. J., & Shope, S. M. (2001). The CERTT-UAV Synthetic Task: Validity, Flexibility, Availability. Paper presented at the Air Force Office of Scientific Research Workshop on Team Performance, October 16-17, Fairfax, VA.
20) Cooke, N. J. (2001). Team Cognition: What Have We Learned? Paper presented at the Air Force Office of Scientific Research Workshop on Team Performance, October 16-17, Fairfax, VA.
21) Cooke, N. J. (2002). Cognitive Task Analysis for Teams. On-line CTA Resource Seminar sponsored by Aptima and Office of Naval Research, October 11, US Positioning, Las Cruces, NM.
22) Cooke, N. J. (2002). Diagnosing Team Performance Through Team Cognition, Paper presented at ONR-NMSU Workshop on New Directions in Cognitive Science, October 25-26, New Mexico State University, Las Cruces, NM.
23) Cooke, N.J. (2003). Assessing Team Cognition. Invited Talk, Los Alamos National Laboratory, June 2, Los Alamos.
24) Cooke, N.J., & Kiekel. P. A. (2003). Team Cognition. AFRL-Rome sponsored seminar on Cognitive Systems Engineering. August 5-7, Hamilton, NY.
25) Cooke, N.J. (2003). Assessing Team Cognition. Invited talk to the Air Force Research Lab. August 28, Mesa, AZ.
26) Cooke, N.J. (2003). Assessing Team Cognition. Invited talks to Arizona State University's Cognition and Behavior Seminar. September 22, 2003.

27) Cooke, N. J. (2003). Playing well with others: Emotional intelligence meets team performance. Paper presented at ETS Workshop: Emotional Intelligence: Knowns and Unknowns, November 15, Princeton, NJ.
28) Cooke, N.J. (2003). Command-and-Control Teams: The Outcome of Cognitive Engineering. Journeys of the Mind Series, November 18, Arizona State University East, Mesa, AZ.
29) Cooke, N. J. (2004). Team cognition in distributed command-and-control. Paper presented at AFOSR Cognitive Decision Making Program Review Workshop, March 9-10, Chandler, AZ.

6.0 GLOSSARY

AFOSR – Air Force Office of Scientific Research
ASU – Arizona State University
AVO – Air Vehicle Operator
AWACS - Airborne Warning and Control System
CAM – Cognitive Abilities Measurement
CERTT Lab - Cognitive Engineering Research on Team Task Laboratory
CIP – Critical Incident Process
Critical Waypoint – A waypoint that has to be visited by UAV team for a successful mission
DEMPC – Data Exploitation, Mission Planning, and Communications Operator
DME – Distributed Mission Environment
DMT – Distributed Mission Training
DURIP – Defense University Research Instrumentation Program
Effective Radius – Area surrounding a waypoint in which airspeed and altitude restrictions are in effect and camera is operable
GDSS – Group Decision Support System
GPA – Grade Point Average
Ground Control – A command-and-control station where operators control UAV systems from the ground
HW – High Workload
IPK - Interpositional Knowledge (knowledge about others' jobs)
JASS – Job Assessment System Software
KNOT – Knowledge Network Organization Tool (Computer Software)
KVM – Keyboard Video Mouse
LW – Low Workload
MCSD - Marlowe-Crowne Social Desirability scale
MTMM – Multi-trait Multi-method
NASA TLX – National Aeronautics and Space Administration Task Load Index
NMSU – New Mexico State University
ONR – Office of Naval Research
PLO – Payload Operator
Pathfinder – Psychological scaling technique used for representing human judgments in graphical form
Predator – Air Force Unmanned Aerial Vehicle
Referent Network – Pathfinder network representing ideal knowledge, generated by experimenters or empirically from expert data
ROZ Entry – Restricted Operating Zone
ROCT – Reserve Officer Training Corps
SA – Situation Awareness
SART - Situation Awareness Rating Technique
SAS – Statistical Analysis Software
SE – Standard Error
STE - Synthetic Task Environment
UAV- Uninhabited Air Vehicle
Waypoint – A named landmark on a map

7.0 APPENDICES

Appendix A

Number of Participants by Organization

Air Force	5
Army	16
Cycling Club	8
Engineering Honor Societies	4
Other Honor Societies	14
Rugby Teams	10
Sororities and Fraternities	3

Appendix B

Components of Revised Individual and Team Performance Scores

Subscore	Subscore Numerator	Subscore Denominator	Transformation	Weight	Relative Weight
AVO					
Alarm Penalty	AVO Alarm Duration	missionTotalSecs	subscore^.5	126.69	4
Warning Penalty	AVO Warning Duration	missionTotalSecs	subscore^.5	25.14	1
Course Dev Penalty	From Flgt_Sum.rds, Sum of all "Sum0fDev"	totalRouteLength	-	287.06	4
AVO Rte Seq Penalty	Planned WPs not Visited** + Visted WPs not Planned - WPs can't make*	total wps planned - WPs can't make*	-	262.94	3
PLO					
Alarm Penalty	PLO Alarm Duration	missionTotalSecs	subscore^.5	567.70	3
Warning Penalty	PLO Warning Duration	missionTotalSecs	subscore^.5	121.96	1
Duplicate Good Photos Penalty	totalGood - totalGoodUnique	film	-	1730.26	4
Missed or Slow Photo Penalty	totalGoodUnique	missionTotalSecs/60	1-subscore	39.02	2
Bad Photo Penalty	Bad Photos	Film	-	178.34	3
DEMPC					
Alarm Penalty	DEMPC Alarm Duration	missionTotalSecs	subscore^.5	265.93	2
Warning Penalty	DEMPC Warning Duration	missionTotalSecs	subscore^.5	30.93	1
Missed CWPs Not Planned Penalty	Critical WPs not planned	unique total wps planned	-	1200.6	4
Alarm WPs Penalty	Hazard/Lost WPs Planned	unique total wps planned	-	692.47	3
Rte Seq Plan Penalty	Rte Seq Plan Violation	total wps planned	-	1177.53	4
TEAM					
Alarm Penalty	TEAM Alarm Duration	missionTotalSecs	subscore^.5	393.22	2
Warning Penalty	TEAM Warning Duration	missionTotalSecs	subscore^.5	112.02	1
Missed or Slow Crit WPs Penalty	critical_reached	missionTotalSecs/60	1-subscore	318.63	3
Missed or Slow Photos Penalty	totalGoodUnique	missionTotalSecs/60	1-subscore	314.96	4

*WPs can't make = total wps planned - the number in the DEMPC route that signifies the last waypoint hit by AVO and planned by DEMPC

** This Planned WPs not visited is not the same number as noted by the rapid file. It is the number of planned WPs not visited out of the unique WPs planne

Appendix C

Critical Incident Process Measure: Low Workload

CRITICAL INCIDENT PROCESS-LW

Use this form for the "Low" Workload scenarios. The non-talking experimenter will evaluate the team process. The following behaviors may or may not occur at the designated event triggers (in italics).

BEGINNING OF MISSION

P1 [very poor/none] [poor] [good] [very good]

Before the team reaches the effective radius of the first target, rate how effectively the TEAM discusses how they will perform during the mission (for example, "good" teams will discuss their plans in a constructive way, perhaps covering the entire mission, "poor" teams may not plan at all, not discuss their performance, or deride themselves and each other).

LVN-OAK ROZ BOX

P2 [none] [poor or unclear] [good, clear information sharing]

Prior to UAV in effective radius of H-AREA or F-AREA or targets within first ROZ box, rate AVO and DEMPC's clarification and acknowledgement of the restrictions and other needed information.

AFTER KGM-FRT CALL-IN

P3 [yes] [no]

Before entering the call-in ROZ box (KGM-FRT), the TEAM explicitly notes and acknowledges the existence of the ROZ's waypoints and targets.

PRK-ASH ROZ BOX

P4 [asked] [did not ask]

Prior to UAV in effective radius of S-STE or MSTE or targets within second ROZ box, PLO asks for PRK-ASH targets before being told by the DEMPC.

KGM-FRT ROZ BOX

P5 [none] [poor or unclear] [good, clear coordination]

Either shortly before entering or while in the third ROZ box (usually KGM-FRT), rate how well AVO and PLO work together to maneuver UAV for photos (this should be evident in their communication). *(Rate the last ROZ box if they don't get to three)*

END OF MISSION

P6 [yes] [no]

Within 5 minutes after end of mission, the TEAM assesses and discusses their performance.

Appendix D

Critical Incident Process Measure: High Workload

CRITICAL INCIDENT PROCESS-HW

Use this form for the High Workload scenarios. The following behaviors may or may not occur at the designated event triggers (in italics).

BEGINNING OF MISSION

P1 [very poor/none] [poor] [good] [very good]

Before the team reaches the effective radius of the first target, rate how effectively the TEAM discusses how they will perform during the mission (for example, "good" teams will discuss their plans in a constructive way, perhaps covering the entire mission, "poor" teams may not plan at all, not discuss their performance, or deride themselves and each other).

LVN-OAK ROZ BOX

P2 [none] [poor or unclear] [good, clear information sharing]

Prior to UAV in effective radius of H-AREA or F-AREA or targets within first ROZ box, rate AVO and DEMPC's clarification and acknowledgement of the restrictions and other needed information.

AFTER PRK-ASH CALL-IN

P3 [yes] [no]

Before entering the call-in ROZ box (PRK-ASH), the TEAM explicitly notes and acknowledges the existence of the ROZ's waypoints and targets.

KGM-FRT ROZ BOX

P4 [asked] [did not ask]

Prior to UAV in effective radius of S-STE or R-STE or targets within third ROZ box, PLO asks for KGM-FRT targets before being told by the DEMPC.

CRT-MNR ROZ BOX

P5 [none] [poor or unclear] [good, clear coordination]

Either shortly before entering or while in the fourth ROZ box (usually CRT-MNR), rate how well AVO and PLO work together to maneuver UAV for photos (this should be evident in their communication). *(Rate the last ROZ box if they don't get to four)*

END OF MISSION

P6 [yes] [no]

Within 5 minutes after end of mission, the TEAM assesses and discusses their performance.

Appendix E

Judgment Process Measure

Good Poor

TALLIES-- **Please tally when you note the following behaviors:**

Communication and Coordination

Examples

- Made clear acknowledgement when an important fact was passed
- Compensated or clarified when a team member performed their job poorly
- Failed to acknowledge when an important fact was passed
- Criticized or did nothing when a team member performed their job poorly

Team decision-making

Examples

- Asserted accurate and critical counter-arguments when making decisions
- Failed to assert/asserted wrong facts
- (e.g. "we can skip this target, because it's not priority")
- Argued logically, or with smooth resolution (esp. at waypoints).
- Bickered or got bogged down by arguing (esp. at waypoints)

Team situation awareness behaviors

Examples

- Team made sure that everyone knew about upcoming targets
- (e.g. stated that a target was approaching AND acknowledged the statement)
- Team got close to a target without clarifying that it was a target.
- Asserted inaccurate information about the immediate situation
- Asserted accurate information about the immediate situation

Please note any other behaviors that were indicative of good or poor team process.

FINAL JUDGEMENT-- Please circle one score for each dimension.

Communication and Coordination

Terrible Poor Average Good Excellent

Team decision-making

Terrible Poor Average Good Excellent

Team situation awareness behaviors

Terrible Poor Average Good Excellent

Process behaviors overall

Terrible Poor Average Good Excellent

Appendix F

Example of Situation Awareness Call-In

Team01 Date ________________ Experimenter ____________________

=-

Mission = 1 Query = 1 [20-25 minutes]

=-

"This is intelligence calling the ____. I have a request for information. Please speak only to EXP so that your responses can be kept secret."

"What is the name of the next target waypoint?"

DEMPC
* Response ________________ * Truth ________________

PLO
* Response ________________ * Truth ________________

AVO
* Response ________________ * Truth ________________

TEAM
* Response ________________ * Truth ________________

=-

Mission = 1 Query = 2 [25-30 minutes]

=-

"This is intelligence calling the ____. I have a request for information. Please speak only to EXP so that your responses can be kept secret."

"How many targets do you think your team will manage to successfully photograph by the end of your 40-minute mission? There are nine targets total."

AVO
* Response ________________ * Truth ________________

PLO
* Response ________________ * Truth ________________

DEMPC
* Response ________________ * Truth ________________

TEAM
* Response ________________ * Truth ________________

Appendix G

Situation Awareness Queries: Experiment 1

Repeated Query:

1. How many targets do you think your team will manage to successfully photograph by the end of your 40 minute mission?

Non-repeated Queries:

2. What is the name of the next target waypoint?
3. What will your altitude be for the next waypoint you will enter?
4. What will your speed be for the next waypoint you will enter?
5. What is the next target waypoint (e.g. boat, building, moose)?
6. What warnings or alarms are currently going off?
7. In the mission so far, have you had to pass through any hazards to get to the current waypoint?
8. What type of waypoint are you heading to now? For example, target, ROZ entry, hazardous, etc.

Situation Awareness Queries: Experiment 2

Repeated Query:

1. How many targets do you think your team will manage to successfully photograph by the end of your 40-minute mission? There are 9 (20) targets total.

Non-Repeated Queries:

2. What type of waypoint are you heading to now? For example, is it a ROZ entry, target, hazard, etc.?
3. What will your speed be for the next waypoint you will enter?
4. What is the next target waypoint a picture of (e.g., boat, building, etc.)?
5. In this mission so far, have you had to pass through any hazards to get to the current waypoint?
6. What will your altitude be for the next waypoint you will enter?

Appendix H

Teamwork Questionnaire

Instructions: You will be reading a mission scenario in which your team will need to achieve some goal. As you go through the scenario in your mind, think about what communications are absolutely necessary among all of the team members in order to achieve the stated goal. For example, does the AVO ever have to call the DEMPC about something? Using checkmarks, indicate on the attached scoring sheet which communications are **absolutely necessary** for your team to achieve the goal.

Scenario: Intelligence calls in a new **priority target** to which you must proceed immediately. There are **speed and altitude restrictions** at the target. You must successfully **photograph the target** in order to move on to the next target. At a minimum, what communications are absolutely necessary in order to accomplish this goal and **be ready to move on to the next target?** (check those that apply)

_____AVO communicates altitude to PLO

_____AVO communicates speed to PLO

_____AVO communicates course heading to PLO

_____AVO communicates altitude to DEMPC

_____AVO communicates speed to DEMPC

_____AVO communicates course heading to DEMPC

_____PLO communicates camera settings to AVO

_____PLO communicates photo results to AVO

_____PLO communicates camera settings to DEMPC

_____PLO communicates photo results to DEMPC

_____DEMPC communicates target name to AVO

_____DEMPC communicates flight restrictions to AVO

_____DEMPC communicates target type (e.g., nuclear plant) to AVO

_____DEMPC communicates target name to PLO

_____DEMPC communicates flight restrictions to PLO

_____DEMPC communicates target type (e.g., nuclear plant) to PLO

Appendix I

Empirical Taskwork Referents

In previous studies, a logical referent network generated by the experimenters served as the key with which taskwork knowledge was evaluated. In Experiment 1, empirical referents were derived for the AVO, PLO, DEMPC, and Team based on the taskwork knowledge networks of the top five performing (determined with the original performance scores) individuals (or teams) over the first three experiments conducted in the UAV-STE. For example, in constructing the AVO empirical referent, we gathered the taskwork networks of the five highest performing AVOs across three experiments (N = 68). The links in the AVO empirical referent reflected the links contained in the majority (i.e., at least three) of the top five performing AVO networks. The team networks, from the top five performing teams, used in constructing the team empirical referent were the teams' holistic networks, which were generated from the taskwork ratings collected at the team level. Alternative approaches to determining the team networks include 1) averaging individual ratings in order to construct a network representative of the team knowledge and 2) using the union of the links in the three individual networks as the team network. We felt that the team networks generated from the holistic ratings were most representative of the teams' knowledge whereas the two alternative approaches did not seem as appropriate for teams with different roles. The basis for deriving new referents empirically stemmed from the notion that experimenters' knowledge of the task is likely more extensive and developed across all roles and thus, may not serve as a proper comparison against participants who are less experienced and knowledgeable of other roles.

The empirically derived referents are listed below in Figures I1 – I4.

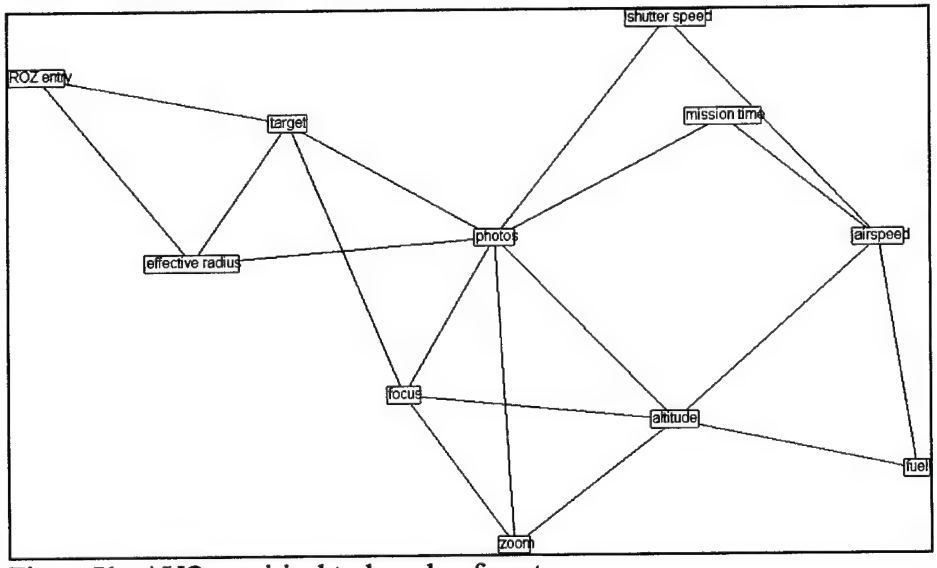


Figure I1. AVO empirical taskwork referent.

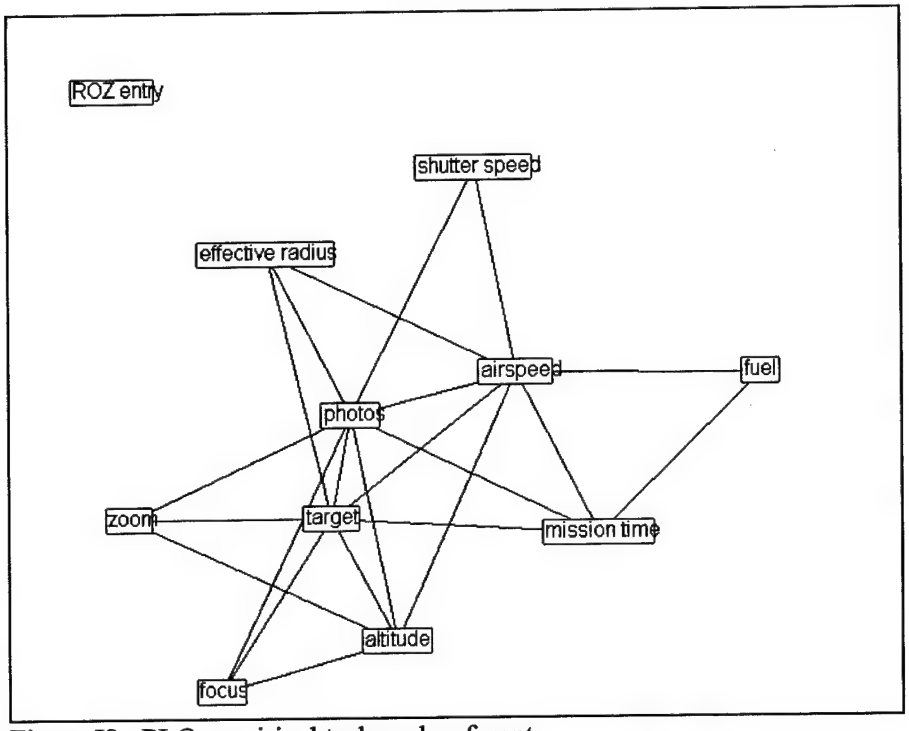


Figure I2. PLO empirical taskwork referent.

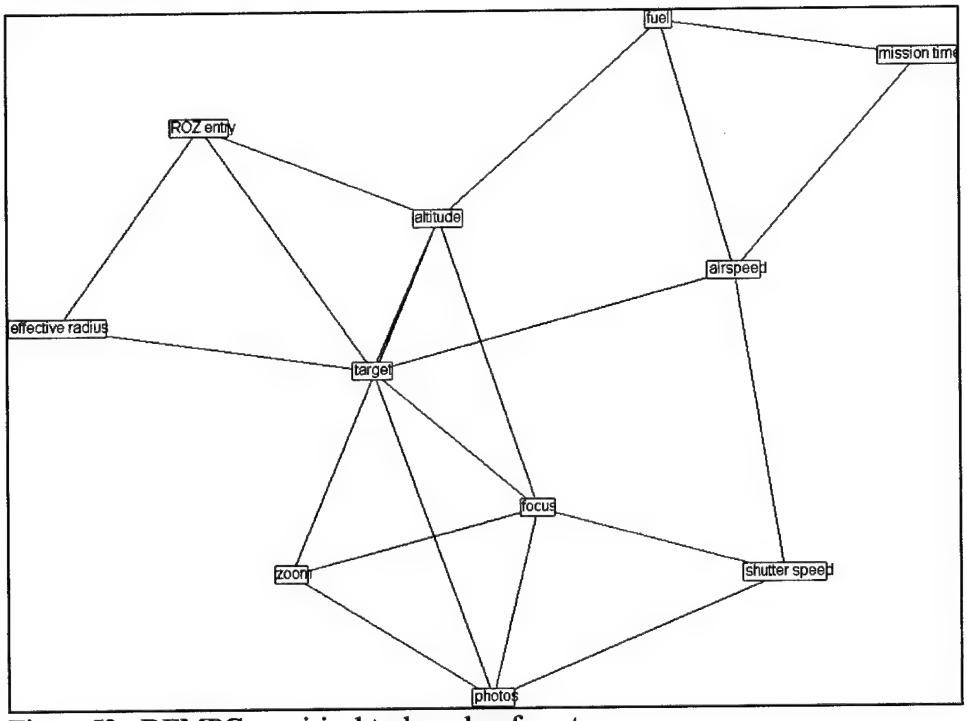


Figure I3. DEMPC empirical taskwork referent.

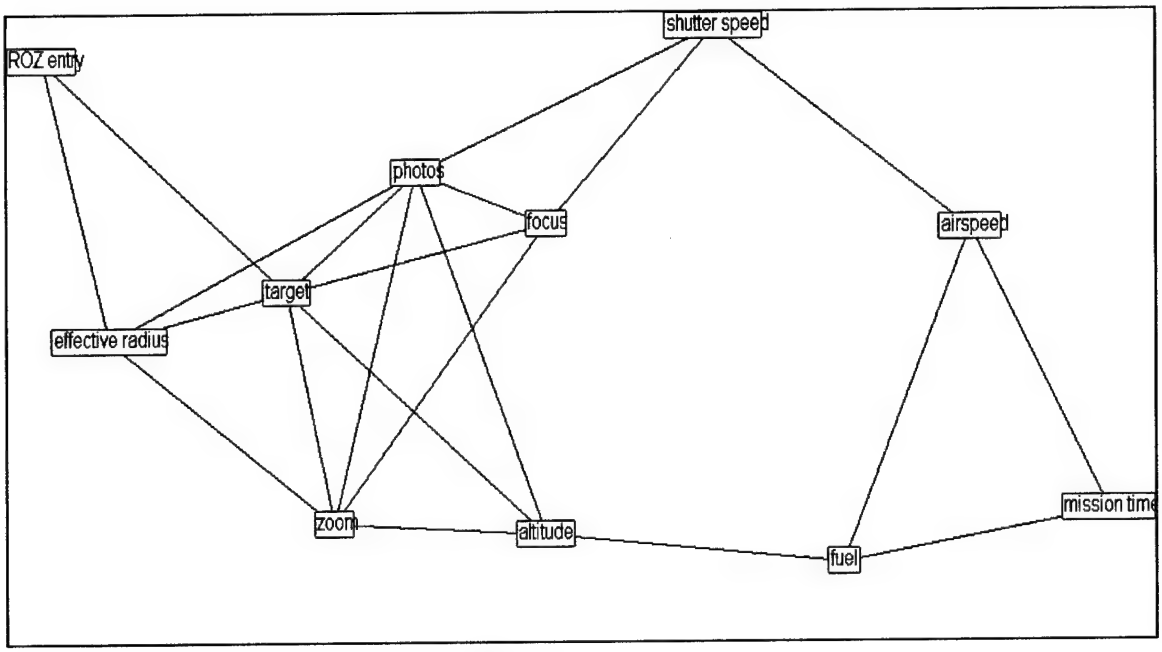


Figure I4. Team empirical taskwork referent.

Appendix J

Secondary Questions

For each team member and for the team as a whole, is knowledge of long-term UAV mission goals (e.g., what makes a successful mission) high, low, or somewhere in-between? (Rate yourself where appropriate. Circle one for each question.)

1. The AVO's knowledge of long-term mission goals is
(low) 1 2 3 4 5 (high)

2. The PLO's knowledge of long-term mission goals is
(low) 1 2 3 4 5 (high)

3. The DEMPC's knowledge of long-term mission goals is
(low) 1 2 3 4 5 (high)

4. The team's knowledge of long-term mission goals is
(low) 1 2 3 4 5 (high)

For each team member and the team as a whole, is knowledge of short-term UAV mission goals (e.g., what needs to be done next) high, low, or somewhere in-between? (Rate yourself where appropriate. Circle one for each question.)

5. The AVO's knowledge of short-term mission goals is
(low) 1 2 3 4 5 (high)

6. The PLO's knowledge of short-term mission goals is
(low) 1 2 3 4 5 (high)

7. The DEMPC's knowledge of short-term mission goals is
(low) 1 2 3 4 5 (high)

8. The team's knowledge of short-term mission goals is
(low) 1 2 3 4 5 (high)

For each team member and the team as a whole, is the ability to request information from the correct person high, low, or somewhere in-between? (Rate yourself where appropriate. Circle one for each question.)

9. The AVO's ability to request information from the correct person is
(low) 1 2 3 4 5 (high)

10. The PLO's ability to request information from the correct person is
(low) 1 2 3 4 5 (high)

11. The DEMPC's ability to request information from the correct person is
(low) 1 2 3 4 5 (high)

12. The team's ability to request information from the correct person is
(low) 1 2 3 4 5 (high)

For each team member and the team as a whole, is the ability to supply the correct information to the correct person high, low, or somewhere in-between? (Rate yourself where appropriate. Circle one for each question.)

13. The AVO's ability to supply correct information to the correct person is
(low) 1 2 3 4 5 (high)

14. The PLO's ability to supply correct information to the correct person is
(low) 1 2 3 4 5 (high)

15. The DEMPC's ability to supply correct information to the correct person is
(low) 1 2 3 4 5 (high)

16. The team's ability to supply correct information to the correct person is
(low) 1 2 3 4 5 (high)

Please select the best possible answer from the four alternatives for each of the following four questions (Circle the letter corresponding to the best answer).

17 . Five miles outside of the effective radius of a target, which of the following activities is least important?

a. DEMPC conveys effective radius to AVO and PLO

b. PLO sets camera in accordance with UAV altitude and speed

c. AVO checks fuel level and sets flaps appropriately

d. DEMPC communicates target type (e.g., nuclear facility) to PLO

18 . Inside of the effective radius of a target, which of these activities would indicate the poorest planning?

a. PLO instructs AVO to be above 3000 feet for appropriate zoom

b. AVO tracks effective radius and turns around if necessary

c. DEMPC communicates target type (e.g., nuclear facility) to PLO

d. PLO takes photos until one is acceptable

19. Imagine that your team must navigate through a hazardous waypoint to reach the next target, which is the least relevant example of requesting information?

a. AVO requests hazardous waypoint name from DEMPC

b. PLO requests an effective radius from DEMPC

c. AVO requests flight restrictions from DEMPC

d. DEMPC requests a course change from AVO

20. Imagine that your team has just photographed a target, which is the least helpful example of supplying information.

a. DEMPC informs AVO about upcoming flight restrictions

b. PLO informs AVO and DEMPC about the status of the picture

c. AVO informs DEMPC about current airspeed

d. DEMPC informs PLO about upcoming flight restrictions

Appendix K

Leadership Survey
(Short Version)

Instructions: The purpose of this survey is to examine leadership roles. Please answer all questions honestly in the context of all 7 missions.

Please rate the extent to which the AVO:

	Never						Always
a) assumed a leadership role	1	2	3	4	5	6	7
b) led the conversation	1	2	3	4	5	6	7
c) influenced group goals and decisions	1	2	3	4	5	6	7

Please rate the extent to which the PLO:

a) assumed a leadership role	1	2	3	4	5	6	7
b) led the conversation	1	2	3	4	5	6	7
c) influenced group goals and decisions	1	2	3	4	5	6	7

Please rate the extent to which the DEMPC:

a) assumed a leadership role	1	2	3	4	5	6	7
b) led the conversation	1	2	3	4	5	6	7
c) influenced group goals and decisions	1	2	3	4	5	6	7

Appendix L

Post-Mission Questions

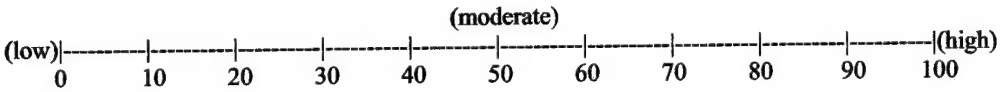
N.A.S.A. T.L.X. (task load index)

Instructions: Since workload is something that is experienced individually by each person, there are no effective 'rulers' that can be used to estimate the workload of different activities. One way to find out about workload is to ask people to describe the feelings they experienced. Because workload may be caused by different factors, we would like you to evaluate several of them individually. The following set of five workload scales was developed for you to use in evaluating your experiences during the mission.

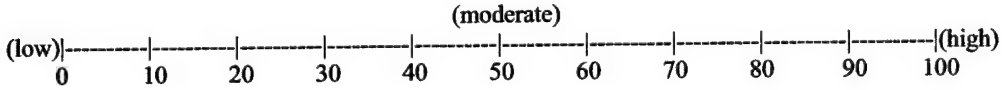
You are going to be responding to 5 different rating scales regarding your last UAV mission. It is important to remember that these scales are subjective. There is no right or wrong answer. Do not spend too much time on any one item. Your' initial feeling is probably the best response.

The following are the definitions of each of the 5 rating scales. Rate each scale after reading through the definition and considering it in the context of your last UAV mission by marking the appropriate location on the ticked line (i.e., low, high, or somewhere in-between).

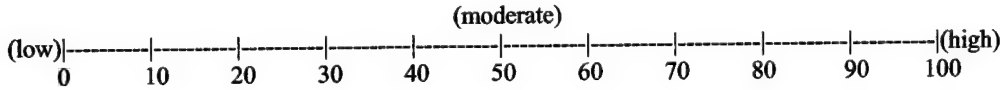
Mental demands: How high or low is the amount of mental and perceptual activity required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)?



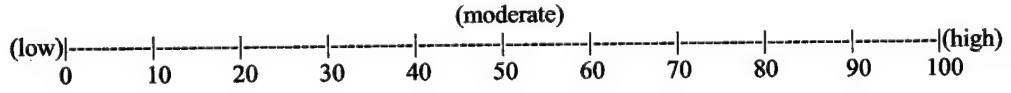
Physical demands: How high or low is the amount of physical activity that is required (e.g., pushing, pulling, turning, controlling, activating, etc.)?



Temporal demands: How high or low is the amount of time pressure you feel due to the rate or pace at which the task (or task elements) occurs?



Performance demands: How important is it to be satisfied with your own performance: Is the importance of your individual score high or low?



Teammate demands: How high or low is the amount of pressure you feel due to demands created by teammates (e.g., need for information, number of communications, etc.)?

(moderate)

(low)|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|(high)

0 10 20 30 40 50 60 70 80 90 100

S.A.R.T. (Situation Awareness Rating Technique)

Instructions: S.A.R.T. uses your own estimates of personal and UAV-dependent factors that may affect your performance and understanding in order to measure situation awareness. In short, we are measuring your ability to apply the meaning of events and elements in the task to mission goals.

You are going to be responding to 14 different rating scales regarding your last UAV mission. It is important to remember that these scales are subjective. There is no right or wrong answer. Do not spend too much time on any one item. Your' initial feeling is probably the best response.

The following are the definitions of each of 14 rating scales. Carefully read through each definition and circle an appropriate number from 1 (low) to 7 (high).

1. **Demand on Cognitive Resources**: During your last mission, were there many difficult situations demanding constant attention and mental effort (high) or was it easy and minimally demanding (low)?

   (low) 1---------2---------3---------4---------5---------6---------7 (high)

2. **Instability of Situations**: During your last mission, were situations likely to change suddenly (high) or were most situations slow with easily predictable outcomes (low)?

   (low) 1---------2---------3---------4---------5---------6---------7 (high)

3. **Complexity of Situations**: During your last mission, were situations complex with many interrelated ideas (high) or were most situations straight-forward (low)?

   (low) 1---------2---------3---------4---------5---------6---------7 (high)

4. **Variability of Situations**: During your last mission, were there a large number of things changing simultaneously (high) or did very few things change simultaneously (low)?

   (low) 1---------2---------3---------4---------5---------6---------7 (high)

5. **Supply of Cognitive Resources**: During your last mission, were you able to pay a lot of attention to problems that arose (high) or did you have a limited amount of attention (low)?

   (low) 1---------2---------3---------4---------5---------6---------7 (high)

6. **Readiness**: During your last mission, were you able to anticipate events and respond quickly (high) or were you hard pressed to keep up with evolving situations (low)?

(low) 1---------2---------3---------4---------5---------6---------7 (high)

7. **Concentration of Attention**: During your last mission, were you always focused on the task at hand (high) or did controls and communication distract you (low)?

(low) 1---------2---------3---------4---------5---------6---------7 (high)

8. **Division of Attention**: During your last mission, were you able to consider current and future events simultaneously (high) or did you focus on only one thing at a time (low)?

(low) 1---------2---------3---------4---------5---------6---------7 (high)

9. **Spare Mental Capacity**: During your last mission, do you think you could have dealt with an additional number of mission elements (high) or did the mission take all your mental capacity (low)?

(low) 1---------2---------3---------4---------5---------6---------7 (high)

10. **Understanding of the Situation**: During your last mission, did you usually have a good understanding of the mission goals (high) or were you uncertain about mission goals (low)?

(low) 1---------2---------3---------4---------5---------6---------7 (high)

11. **Information Quantity**: During your last mission, did you receive a great deal of useful information (high) or was very little of the information of much use to you (low)?

(low) 1---------2---------3---------4---------5---------6---------7 (high)

12. **Information Quality**: During your last mission, was the information communicated to you accurate and precise (high) or was the information noisy with high levels of uncertainty (low)?

(low) 1---------2---------3---------4---------5---------6---------7 (high)

13. **Familiarity with Environment**: During your last mission, did you have a great deal of relevant experience (high) or did you find significant aspects of the mission unfamiliar to you (low)?

(low) 1---------2---------3---------4---------5---------6---------7 (high)

14. **Situation Awareness**: During your last mission, did you have a complete picture of how various elements would affect the mission and could you anticipate mission-critical events and decisions (high) or did you have limited ability to predict the impact of on-going activity on future events and overall mission goals (low)?

(low) 1---------2---------3---------4---------5---------6---------7 (high)

Appendix M

Experiment 1 Debriefing Questions

Demographic Questions

1. Rank
2. Major
3. Aviation Experience
4. Ethnicity
5. Class (e.g., freshman)
6. Gender
7. GPA

Miscellaneous Questions (Scale: 0-disagree to 4-agree)

8. I enjoyed participating in this study
9. I enjoyed the team task part of this study
10. I would welcome the opportunity to participate in this study in the future
11. I would like to work with my fellow team members again
12. I like playing video and computer games
13. I like to be part of a team
14. I was a successful member of the team
15. My team worked well together
16. I performed well on this task
17. The AVO was competent
18. The AVO contributed to the team
19. The AVO tried hard
20. The AVO was lucky
21. The AVO had an easy task
22. The AVO was likable
23. The PLO was competent
24. The PLO contributed to the team
25. The PLO tried hard
26. The PLO was lucky
27. The PLO had an easy task
28. The PLO was likable
29. The DEMPC was competent
30. The DEMPC contributed to the team
31. The DEMPC tried hard
32. The DEMPC was lucky
33. The DEMPC had an easy task
34. The DEMPC was likable
35. My team performed well on this task
36. My individual performance is important to our team
37. Performance was evaluated at the individual level
38. Performance was evaluated at the team level

Trust Questions (Scale: 0-disagree to 4-agree)

39. My teammates can be counted on to do what they say they will do
40. Some teammates hold back information that is critical to the mission and our performance
41. I trust the AVO (If you were the AVO, rate yourself)
42. I trust the PLO (If you were the PLO, rate yourself)
43. I trust the DEMPC (If you were the DEMPC, rate yourself)

Anxiety Questions (Scale: 0-disagree to 4-agree)

44. I rarely worry about seeming foolish to others
45. I am often afraid that I may look ridiculous or make a fool of myself
46. If someone is evaluating me, I tend to expect the worst

Team Member Prior Familiarity (Scale: 0-Me, 1-Stranger to Me, 2-Somewhat Familiar, and 3-Well Known to Me)

47. AVO Familiarity
48. PLO Familiarity
49. DEMPC Familiarity

Social-Desirability Questions (True/False)

50. Before voting, I thoroughly investigate the qualifications of all the candidates
51. I never hesitate to go out of my way to help someone in trouble
52. It is sometimes hard for me to go on with my work if I am not encouraged
53. I have never intensely disliked anyone
54. On occasion I have had doubts about my ability to succeed in life
55. I sometimes feel resentful when I don't get my way
56. I am always careful about my manner of dress
57. My table manners at home are as good as when I eat out in a restaurant
58. If I could get into a movie without paying and be sure I was not seen, I would probably do it
59. On a few occasions, I have given up doing something because I have thought too little of my ability
60. I like to gossip at times
61. There have been times when I felt like rebelling against people in authority even though I knew they were right
62. No matter who I'm talking to, I'm always a good listener
63. I can remember 'playing sick' to get out of something
64. There have been occasions when I took advantage of someone
65. I'm always willing to admit when I make a mistake
66. I always try to practice what I preach
67. I don't find it particularly difficult to get along with loud-mouthed, obnoxious people
68. I sometimes try to get even, rather than forgive and forget
69. When I don't know something I don't at all mind admitting it
70. I am always courteous, even to people who are disagreeable
71. At times I have really insisted on having things my own way
72. There have been occasions when I felt like smashing things

73. I would never think of letting someone else be punished for my wrongdoings
74. I never resent being asked to return a favor
75. I have never been irked when people expressed ideas very different from my own
76. I never make a long trip without checking the safety of my car
77. There have been times when I was quite jealous of the good fortune of others
78. I have almost never felt the urge to tell someone off
79. I am sometimes irritated by people who ask me to do favors
80. I have never felt that I was punished without cause
81. I sometimes think when people have a misfortune they only got what they deserved
82. I have never deliberately said something that hurt someone's feelings

Team Strategy Question (Open-ended)
Describe the strategies that you and your fellow team members used to generate conceptual relatedness ratings at the team level, given discrepancies among two or three individual ratings for that concept pair.

Appendix N

Experiment 2 Debriefing Questions

Demographic Questions

1. Rank
2. Major
3. Aviation Experience
4. Ethnicity
5. Class (e.g., freshman)
6. Gender
7. GPA
8. Age

Team Member Prior Familiarity (Scale: 0-Me, 1-Stranger to Me, 2-Somewhat Familiar, and 3-Well Known to Me)

9. AVO Familiarity
10. PLO Familiarity
11. DEMPC Familiarity

Appendix O

Demographics Questionnaire

Please answer the following questions. Answer N/A if the question is not applicable to you.

1. If you are in the military, what is your rank?

2. If applicable, what level of aviation training have you accomplished?

3. What is your ethnicity
   a. African American
   b. Asian American
   c. Caucasian
   d. Hispanic
   e. Native American
   f. Other

4. If you're a student, what is your major?

5. If you're a student, what class are you in?
   a. Freshman
   b. Sophomore
   c. Junior
   d. Senior
   e. Graduate Student

6. What is your gender?
   a. Male
   b. Female

7. What is your age?

Appendix P

Debriefing Interview Form

EXPERIMENTER INSTRUCTIONS: Ask each question to each individual separately, so that each individual's responses are not heard by the other participants. For questions that include a scale, try to answer the question using the scale, but feel free to write additional comments in the space provided.

1. Describe your previous interactions with the other team members. Were you part of a team or a regular group? What was the task your team/group performed? Explain

2. What were the characteristics of the team task you and your teammates performed? Circle all that apply.
   a. It involved planning
   b. It involved sharing information
   c. It involved decision making
   d. It involved verbal communication
   e. It involved computer-mediated communication
   f. It involved coordination
   g. It involved physical activities
   h. Sometimes periods of the task were busier, or required more resources than other parts of the task
   i. One of us was designated as the leader
   j. A leader wasn't pre-designated, but one of us usually acted as the leader
   k. We were geographically dispersed
   l. We used technology to communicate
   m. Other?

   Additional Comments:

3. How experienced (i.e., how much of an expert) are you at the team task described above?
   a. Novice (very inexperienced)
   b. Somewhat experienced
   c. Fairly experienced
   d. Expert (very experienced)

   Additional Comments:
4. How long have you worked together as a team?
   a. A day or less

b. More than a day but less than a week
c. Between a week and a month
d. 2-6 months
e. 7-12 months
f. 1-5 years
g. More than 5 years

Additional Comments:

5. During the period of time that you have worked together as a team, how frequently did you work together?
   a. Constantly
   b. Occasionally
   c. Seldom
   d. Very rarely

   Additional Comments:

6. Did all 3 of you have the same job on this team? For example, did all of you perform the same exact task?
   a. Yes, we all did the same thing
   b. No, we all played different roles in the team
   c. Two of us had the same role and the third person had a distinct role
   d. We all played different roles, but we knew each other's roles and could substitute for one another if necessary

   Additional Comments:

7. What mode of communication did you use as a team? Give a percentage out of 100 of how much of the time you used each communication mode.

   a. Face-to-face _____
   b. Over the phone _____
   c. Over the internet _____
   d. Other Computer mediated _____
   e. Other _____

   Additional Comments:

8. How well do you know AVO?
   a. This is the AVO
   b. Not very well
   c. Slightly well
   d. Moderately well
   e. Very well

9. How well do you know PLO?
    a. This is the PLO
    b. Not very well
    c. Slightly well
    d. Moderately well
    e. Very well

10. How well do you know the DEMPC?
    a. This is the DEMPC
    b. Not very well
    c. Slightly well
    d. Moderately well
    e. Very well

In the future, we may have more questions for you about your experience working in this team. If you would like to give us permission to contact you, please sign below and fill in your contact information.

Name: _____

Signature: _____ Date:_____

Phone:_____

E-mail:_____

Appendix Q

Questions Appended to Debriefing Interview Form for UAV Team

Additional Questions for UAV Team:

11. How is our task different from the UAV tasks you regularly perform?

12. What type of UAVs do you have experience with? How are they different from this?

13. Was the team interaction required in this study similar to the team interaction required in your regular job of working with UAVs?

14. Does it seem that the team interaction requirements in our study were exaggerated?

15. Is there any team interaction missing from our task here that is normally required of your UAV team?

16. Did your past experience help you in performing our UAV task?

17. Is there anything about your previous team coordination and interaction that helped or hurt you in our UAV task?

18. What do you think would happen if you were asked to come back in 3 months in terms of your performance? Would you be able to pick up where you left off in terms of knowing your role and performing similar to how you performed today?

19. What if when you came back in 3 months you were paired with 2 new team members whom which you were not familiar?

Appendix R

Basic Skills Checklist

Have the following behaviors performed by the three team members in order and check them off as they are accomplished. With two experimenters, the DEMPC and AVO checks can be conducted in parallel with the PLO checks following

COMMUNICATION CHECKS

Everyone should put headsets on, including the experimenters. Experimenters talk to team members over the headsets conducting the following checks. Adjust microphones and instruct on push-to-talk button and intercom as needed.

Experimenter queries each team member in turn:

___ Experimenter can hear AVO
___ AVO can hear Experimenter
___ Experimenter can hear PLO
___ PLO can hear experimenter
___ Experimenter can hear DEMPC
___ DEMPC can hear experimenter

Experimenter queries each team member in turn:
___ Experimenter can hear everyone
___ AVO can hear PLO and DEMPC
___ PLO can hear AVO and DEMPC
___ DEMPC can hear AVO and PLO

Instruct team members to push appropriate button to talk.
___ AVO can talk to DEMPC only
___ PLO can talk to AVO only
___ DEMPC can talk to PLO only

Remove and stow headsets. Start the UAV simulation (Training Mission- see Manual Section V). Ask the team members to do each of the following activities and check them off as they are observed. In both conditions, the participants should stay glued to their stations.

DEMPC CHECKS

"As the Dempc, your job is to plan the UAV flight route. This is the initial route given to you by Intel. Every waypoint on this list corresponds to a point on your world map. You need to look through your list and identify all the necessary waypoints for your mission, such as ROZ entry/exits and targets. You also need to remove possible hazards and unnecessary waypoints. You want to get five waypoints that you plan to attend in a row so you can sequence them and send the route to the AVO. Remember, once you hit sequence

you cannot change any of the five waypoints that are highlighted. Start at the top of the list and identify the waypoints listed by running the cursor over the corresponding point on the map. All necessary waypoint information is found in your information window." [have Dempc do this until they reach BEB].

___ Delete waypoint BEB from the flight plan:
"Since BEB is a hazard you need to remove that point from your list."
[ask if they remember how to delete a waypoint and show them if they need help]

___ Insert waypoint BYU into the flight plan between MON and WIC
"BYU is a ROZ entry that's not listed in your initial route list. You must go through a ROZ entry before you take pictures of any targets within a ROZ box so you need to add this waypoint."
[ask if they remember how to insert a waypoint and show them if they need help]

___Identify the effective radius of BYU
"Part of your job is to communicate all necessary information about waypoints to your team members, such as airspeed or altitude restrictions and the effective radius. Remember, as long as a waypoint has restrictions you will receive a hazard warning. You want to encourage your team to get through those waypoints as quickly as possible."
[ask dempc to identify the effective radius]

___ Sequence the plan until the following subset of 5 is highlighted: MAR, SAN, TKE, MON, BYU.
"Once you have five good waypoints you can hit the sequence button. Notice that once you sequence the route it shows up as a line on your world map."
[help the dempc get the above five waypoint sequenced]

___ Send this route
"Now that your waypoints are sequenced you can send this route to the AVO"
[have dempc hit send route button]

AVO CHECKS

"As the AVO, your job is to fly the UAV. The first thing you need is the route from the Dempc. You can ask for this by hitting the request flight plan button or by verbally asking the DEMPC. Once the Dempc sends you the route it will show up on the moving map. Notice that the first waypoint on the map is MAR. You need to enter this point in the box labeled 'To Waypoint'." [ask if AVO remembers how to cue a waypoint and put it into the 'To waypoint' box. If not show them how].

___ Adjust course so that you are heading to the "To Waypoint," MAR. Keep adjusting course throughout checks to minimize deviation.

"Once you have a waypoint in the 'to' box, the 'to goal' box will give you information on the bearing you need to set, the time and distance to the target, and your course deviation. You want to keep the deviation as low as possible."
[ask the AVO if they remember how to adjust the course and if not show them]

___ Change the queued waypoint to SAN.
"It is a good idea to have the queued waypoint ready to go. The next waypoint on your moving map is SAN."
[ask the AVO if they remember how to que the waypoint and if not show them]

___ Adjust airspeed between 100 & 200
"Most of your waypoints will have restrictions on airspeed and altitude. You may need to get this information from the DEMPC."
[have AVO ask dempc for restrictions and make sure the write them down. Ask if they remember how to adjust airspeed and if not show them.]

___ Adjust altitude between 500 & 1000
[ask the AVO if they remember how adjust altitude and if not show them]

___ Raise & lower flaps and landing gear
"You may need to adjust your flaps and landing gear. Your landing gear and flaps should be UP when your flying ABOVE 4000 ft. or you will slow the UAV. Gear and flaps should be DOWN when you're BELOW 1000 ft."
[have the AVO practice raising and lowering the flaps and landing gear]

___ Make SAN the new "To Waypoint"
"Once you are within the effective radius of MAR you can change the 'to waypoint' to SAN."
[ask AVO to change 'to waypoint']

___ Adjust course to head toward SAN. Keep adjusting course throughout checks to minimize deviation.

___ Make sure AVO knows where to find Refuel button on the left side of the workstation.
"You need to keep an eye on your fuel."
[ask AVO if they remember how to refuel and if not show them]

___ The effective radius for SAN is 5. What does this mean?
[make sure the AVO can tell you about the effective radius and if they don't understand then explain]

Keep adjusting course to head toward SAN maintaining current airspeed and altitude. This is necessary for the PLO checks.

PLO CHECKS

"As the PLO, your job is to take pictures of targets. You may need to get information on upcoming targets from your team members."

___ The upcoming waypoint SAN is a target. The effective radius is 5 miles. Find the photo requirements for this target.
"You need to scroll through the alphabetical target list until you find the waypoint. Called in targets are not listed but you can hit the current button and this will give you settings for the waypoint in the 'to waypoint' box.
[make sure the PLO knows how to get the required settings and if not show them.]

___ Set the camera settings.
"The camera settings need to be accurate in order for the picture to be good. They type of camera you need is given in your required settings. The shutter speed and focus are based on the UAVs current airspeed and altitude settings. You will need to confirm these with the AVO. [have them refer to the cheat sheets to set properly]. **The apperature is based on the light meter found on your second screen. The zoom is given in the required settings. Remember zoom x1 requires an altitude of 3000 ft or less and zoom x10 requires an altitude of 3000 feet or more. You may need to work with the AVO to get the altitude you need to take the picture."**
[make sure the PLO double checks to make sure all settings correct]

___ The effective radius for SAN is 5. What does this mean?
[makes sure PLO tells you that they need to be in effective radius to take picture]

___ Take a picture. If it is good press accept. If it's not keep adjusting settings until it is.
"Once you are in the effective radius you can take a picture. You can check the quality of your picture against other pictures in the book at you station. Once you take a good picture remember to hit the accept button otherwise you will not get credit for the picture."
[have PLO keep taking picture until it is good]

___ Make sure PLO knows where to find Battery, Temperature, Lens and Film buttons on the left hand side of the workstation.
"If you have a warning the "take picture" button will turn red and you will not be able to take a photo. Also, remember that the UAV must be steady to take a picture. If the AVO is changing course, airspeed or altitude your "take picture" button will be red."

Appendix S

Effects of Increased Workload on Team Performance and Subjective Estimates of Workload

Effect of Workload on Primary Task Performance versus Secondary Task Performance

Here we describe analyses that were done in order to examine the effect of workload on dual task performance in Experiments 1 and 2, where dual task performance is defined in terms of its effects on primary task performance and secondary task performance. A primary task is defined as the target of evaluation, or the task whose priority is emphasized (Wikens & Hollands, 2000). In contrast, a secondary task is less important and should be carried out with resources not allocated to the primary task. In our UAV mission, the secondary task is composed of the sub-tasks of attending to warnings and alarms. The primary task incorporates all other sub-tasks (e.g., taking photos, planning the route, etc.).

Primary and secondary task performance scores were calculated by subtracting points for penalties on particular components of the mission. For example, at the team level, primary task performance penalties can be incurred if the team fails to photograph necessary targets or if critical waypoints are not visited. Each component was weighted according to its importance to the mission goals. Furthermore, primary task performance and secondary task performance were standardized. The data presented here represent performance penalties in which high, positive numbers reflect a higher penalty (i.e., poorer performance) and low or negative numbers reflect low penalty (i.e., better performance).

The analysis of dual task performance was performed at the team level as well as at each individual level. Primary task performance penalty components are different for the team, AVO, PLO, and DEMPC (see Table S1). That is, the components on which each individual and the team can receive penalties are unique to each individual role and the team as a whole. However, secondary task performance penalty components include: (1) time spent in warning state, and (2) time spent in alarm state for all individuals as well as the team, though the specific alarm and events differed for each role.

Table S1
Primary and Secondary Task Performance Penalty Components at the Team Level and Each Individual Level

	Primary Task Performance Penalty Components		
Team	Failure to visit critical waypoints	Missed photos	
AVO	Course deviation	Route deviation	
PLO	Duplication of good photos	Missed photos	Bad photos
DEMPC	Failure to plan critical waypoints	Planning hazardous waypoints	Violations of route sequencing
	Secondary Task Performance Penalty Components		
Team	Time spent in alarm state	Time spent in warning state	
AVO	Time spent in alarm state	Time spent in warning state	
PLO	Time spent in alarm state	Time spent in warning state	
DEMPC	Time spent in alarm state	Time spent in warning state	

Each of the four analyses (Team, AVO, PLO, DEMPC) examined the effects of workload (a repeated measure; low vs. high) and dispersion (a between-subjects factor; co-located vs.

distributed) on dual task performance using a doubly multivariate analysis of variance (MANOVA). Mission 4 data were used as an estimate for low workload and Mission 5 data were used as an estimate for high workload.

Experiment 1. Table S2 shows the *F*-statistics and significance of each test of the MANOVA performed at the team level for Experiment 1. Only significant effects are discussed. There was a significant interaction between dispersion and task (primary vs. secondary) indicating that the effect of dispersion on dual-task performance depended on whether the penalties were from the primary task or secondary task. However, *post hoc* tests did not reveal any significant differences between co-located and distributed teams on (1) primary task performance, $F(1, 18) < 1$, or (2) secondary task performance, $F(1, 18) < 1$. The pattern of effects of dispersion on dual-task performance was reversed depending on the task. Distributed teams received more penalties than co-located teams on the primary task, but co-located teams acquired more penalties than distributed teams on the secondary task. This directional difference between co-located and distributed underlies the significant interaction effect.

Table S2
Primary and Secondary Tasks Performance Penalties in Co-located and Distributed Conditions during Low and High Workload at the Team Level for Experiment 1

	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Mission 4 (LW)	-1.11	-.86	.68	.93	-2.19	-2.51	-.14	.17
Mission 5 (HW)	.27	.12	.54	.58	-.45	-.84	1.11	.95
Mission 4 (LW)	-.61	-.35	.96	.62	-2.09	-1.30	1.08	.56
Mission 5 (HW)	.12	-.50	.78	.91	-1.51	-1.95	1.17	.92

Table S3
F-values from the MANOVA on Team Dual Task Performance for Experiment 1

Effect	*F*
Task * Dispersion	6.06*
Task * Workload	12.25**
Dispersion * Workload	.30
Task * Workload * Dispersion	.86
Dispersion	.11
Workload	.08

df = 1,18 *$p < .05$ **$p < .01$

A significant interaction also emerged between workload and task. A *post hoc* test indicated that the source of the interaction between task and workload stemmed from the fact that primary task performance penalties significantly increased during high workload, $F(1, 18) = 28.14$, $p < .01$, whereas secondary task performance penalties did not significantly change when the high workload mission was introduced, $F(1, 18) = 1.21$.

Table S4 shows the mean, standard deviation, minimum, and maximum of the primary task performance penalty and secondary task performance penalty for co-located and distributed AVOs during low workload and high workload.

Table S4
Primary and Secondary Tasks Performance Penalties in Co-located and Distributed Conditions during Low and High Workload for AVOs for Experiment 1

	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Mission 4 (LW)	-.70	.01	.59	1.18	-1.28	-1.14	.73	2.41
Mission 5 (HW)	-.33	.06	.52	1.12	-1.24	-1.21	.44	2.21
Mission 4 (LW)	-.44	-.21	.92	.85	-1.97	-1.24	1.49	1.04
Mission 5 (HW)	.32	-.34	.81	.98	-1.07	-1.61	1.58	1.49

As with the team-level, an interaction between dispersion and task was found (see Table S5 for *F*-statistics), suggesting that for AVOs the effect of dispersion on dual-task performance depended on whether the penalties were associated with the primary task or secondary task. Again, although *post hoc* tests did not reveal any significant differences between co-located and distributed AVOs on (1) primary task performance, $F(1, 18) = 2.35$, or (2) secondary task performance, $F(1, 18) < 1$, the pattern of effects of dispersion on dual-task performance was reversed depending on the task. Similar to the team level, distributed AVOs received more penalties than co-located AVOs on the primary task but co-located AVOs acquired more penalties than distributed AVOs on the secondary task. This directional difference between co-located and distributed underlies the significant interaction effect. No other effects were significant.

Table S5
F-values from the MANOVA on AVO Dual Task Performance for Experiment 1

Effect	*F*
Task * Dispersion	3.38*
Task * Workload	.17
Dispersion * Workload	2.42
Task * Workload * Dispersion	1.41
Dispersion	.51
Workload	.09

df = 1,18 *$p < .10$

Table S6 shows the mean, standard deviation, minimum, and maximum of the primary task performance penalty and secondary task performance penalty for co-located and distributed PLOs during low workload and high workload.

Table S6
Standardized Primary and Secondary Tasks Performance Penalties in Co-located and Distributed Conditions during Low and High Workload for PLOs for Experiment 1

	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Mission 4 (LW)	-.49	-.42	.36	.55	-.90	-.94	.14	.77
Mission 5 (HW)	-.21	.01	.59	1.36	-.89	-.91	.86	3.71
Mission 4 (LW)	-.36	-.76	.58	.56	-1.17	-1.39	.57	.39
Mission 5 (HW)	.09	.06	.74	.54	-.69	-.62	1.44	.94

No significant effects emerged (see Table S7 for *F*-statistics).

Table S7
F-values from the MANOVA on PLO Dual Task Performance for Experiment 1

Effect	*F*
Task * Dispersion	.87
Task * Workload	1.26
Dispersion * Workload	2.61
Task * Workload * Dispersion	.17
Dispersion	.02
Workload	.10

df = 1,18

Table S8 shows the mean, standard deviation, minimum, and maximum of the primary task performance penalty and secondary task performance penalty for co-located and distributed DEMPCs during low workload and high workload.

Table S8
Standardized Primary and Secondary Tasks Performance Penalties in Co-located and Distributed Conditions during Low and High Workload for DEMPCs for Experiment 1

	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Mission 4 (LW)	-.96	-.13	.34	1.26	-1.30	-1.21	-.20	2.88
Mission 5 (HW)	.70	1.03	.54	.86	-.01	.27	1.55	2.79
Mission 4 (LW)	-.28	-.09	.77	.73	-.66	-.69	1.80	1.29
Mission 5 (HW)	-.04	-.21	.78	.72	-.67	-.73	1.41	1.12

A significant interaction was found between task and workload (see Table S9 for *F*-statistics and significance), where the effect of workload on dual task performance was moderated by whether the performance penalty was from the primary or secondary task. A *post-hoc* test revealed that, similar to the team level, DEMPCs accrued significantly more primary task performance

penalties in high workload than in low workload, $F(1, 18) = 28.94$, $p < .01$, while their secondary task performance penalties for the most part remained constant across the levels of workload, $F(1, 18) < 1$.

There was also a significant interaction between dispersion and task. A *post-hoc* MANOVA was performed to isolate the source of this interaction. The results indicated that for the primary task, distributed DEMPCs received significantly more penalties than co-located DEMPCs, $F(1, 18) = 3.02$, $p = .10$; however, on the secondary task of monitoring alarms and warnings, there was no difference in penalties received between co-located and distributed DEMPCs, $F(1, 18) < 1$.

Finally, a main effect of workload emerged, where DEMPCs suffered more performance penalties in Mission 5 than in Mission 4 (see Table S9).

Table S9
F-values from the MANOVA on DEMPC Dual Task Performance for Experiment 1

Effect	*F*
Task * Dispersion	4.40*
Task * Workload	34.91***
Dispersion * Workload	2.46
Task * Workload * Dispersion	.12
Dispersion	1.26
Workload	3.13*

df = 1,18 *$p < .10$ **$p < .05$ ***$p<.01$

Experiment 2. Table S10 shows the mean, standard deviation, minimum, and maximum of the primary task and secondary task performance penalties at the team level for co-located and distributed teams during low workload and high workload. Table S10 shows the F-statistics and significance of each test of the MANOVA performed at the team level.

Table S10
Primary and Secondary Task Performance Penalties in Co-located and Distributed Conditions during Low and High Workload at the Team Level in Experiment 2

	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Mission 4 (LW)	-.74	-.84	.98	.67	-1.62	-1.70	1.39	.68
Mission 5 (HW)	.05	.29	.67	.45	-.10	-.27	1.26	.91
Mission 4 (LW)	.03	-.17	1.50	.92	-2.02	-1.64	2.47	1.18
Mission 5 (HW)	.12	.27	.93	.93	-1.11	-1.28	1.51	1.49

Table S11
F-values from the MANOVA on Team Dual Task Performance in Experiment 2

Effect	*F*
Task * Dispersion	0.03
Task * Workload	10.22*
Dispersion * Workload	2.12
Task * Workload * Dispersion	.00
Dispersion	.00
Workload	27.95*

df = 1, 18 *$p < .01$

Only significant effects are discussed. As can be seen a significant main effect of workload emerged, suggesting that teams suffered fewer performance penalties during low workload. There was also a significant interaction between task and workload. A *post hoc* test indicated that the source of the interaction between task and workload resulted from the fact that primary task performance penalties significantly increased during high workload, $F(1, 18) = 18.84, p <.01$, whereas secondary task performance penalties did not significantly change when the high workload mission was introduced, $F(1, 18) < 1$.

Table S12 shows the mean, standard deviation, minimum, and maximum of the primary task performance penalty and secondary task performance penalty for co-located and distributed AVOs during low workload and high workload. Table S13 shows the F-statistics of each test of the MANOVA performed for AVOs. All tests of effects and interactions failed to reach significance at the .10 level.

Table S12
Primary and Secondary Tasks Performance Penalties in Co-located and Distributed Conditions during Low and High Workload for AVOs in Experiment 2

	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Mission 4 (LW)	-.21	-.28	.33	.22	-.54	-.50	.56	.16
Mission 5 (HW)	-.08	-.21	.50	.24	-.51	-.47	1.09	.22
Mission 4 (LW)	.15	.01	1.49	.94	-1.89	-1.47	2.02	1.24
Mission 5 (HW)	.14	.36	.86	1.02	-1.00	-1.53	1.16	1.96

Table S13
F-values from the MANOVA on AVO Dual Task Performance in Experiment 2

Effect	*F*
Task * Dispersion	.56
Task * Workload	.95
Dispersion * Workload	.00
Task * Workload * Dispersion	.55
Dispersion	.49
Workload	1.44

df = 1, 18

Table S14 shows the mean, standard deviation, minimum, and maximum of the primary task performance penalty and secondary task performance penalty for co-located and distributed PLOs during low workload and high workload.

Table S14
Primary and Secondary Tasks Performance Penalties in Co-located and Distributed Conditions during Low and High Workload for PLOs in Experiment 2

	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Mission 4 (LW)	-.29	-.33	.59	.88	-.89	-.88	.61	2.05
Mission 5 (HW)	-.24	.05	.50	.80	-.89	-.70	.89	1.73
Mission 4 (LW)	-.22	-.49	.77	1.00	-1.37	-1.30	.63	1.15
Mission 5 (HW)	.36	.25	.79	.69	-.49	-.56	2.05	1.91

As can be seen from Table S15, there was a significant main effect of workload, where PLOs received more performance penalties under high workload. There was also a significant interaction between task and workload. A *post hoc* test indicated that the source of the interaction between task and workload resulted from the fact that secondary task performance penalties significantly increased during high workload, $F(1, 18) = 6.24$, $p = .02$, whereas primary task performance penalties did not significantly change when the high workload mission was introduced, $F(1, 18) < 1$.

Table S15
F-values from the MANOVA on PLO Dual Task Performance in Experiment 2

Effect	*F*	*p*-Value
Task * Dispersion	.53	.47
Task * Workload	3.15	.09*
Dispersion * Workload	1.41	.25
Task * Workload * Dispersion	.21	.65
Dispersion	.01	.93
Workload	21.82	< .01*

df = 1, 18 *p ≤ .10

Table S16 shows the mean, standard deviation, minimum, and maximum of the primary task performance penalty and secondary task performance penalty for co-located and distributed DEMPCs during low workload and high workload.

Table S16

Primary and Secondary Tasks Performance Penalties in Co-located and Distributed Conditions during Low and High Workload for DEMPCs in Experiment 2

	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Mission 4 (LW)	-.43	-.54	.74	.36	-.86	-.86	1.58	.16
Mission 5 (HW)	-.39	1.26	1.24	.74	-.34	.38	4.17	2.53
Mission 4 (LW)	-.11	-.26	1.05	.59	-.70	-.63	2.73	.97
Mission 5 (HW)	.003	-.12	1.12	.99	-.66	-.66	2.53	1.99

Table S17

F-values from the MANOVA on DEMPC Dual Task Performance in Experiment 2

Effect	*F*
Task * Dispersion	.00
Task * Workload	26. 57*
Dispersion * Workload	.00
Task * Workload * Dispersion	.00
Dispersion	.00
Workload	53. 38*

df = 1, 18 *p < .01

As can be seen from the previous table, a main effect of workload emerged, suggesting that DEMPCs suffered more performance penalties under high workload than under low workload conditions. There was also a significant interaction between task and workload, where the effect of workload on dual task performance was moderated by whether the performance penalty was from the primary or secondary task. A *post-hoc* test revealed that, similar to the team level, DEMPCs accrued significantly more primary task performance penalties in high workload than in low workload, $F(1, 18) = 3.67$, $p = .06$, while their secondary task performance penalties for the most part remained constant across the levels of workload, $F(1, 18) < 1$.

Summary. Results are fairly consistent across both experiments. In general, increased workload resulted in more penalty points on the primary task, with little noticeable secondary task effect. This was true at the team level and for DEMPCs in both experiments. However for the PLOs in Experiment 2 the opposite pattern emerged in which the secondary task, but not the primary task was negatively affected by the change in workload. In Experiment 1, but not Experiment 2 there was also a significant interaction between dispersion and task for teams and AVOs in which distributed teams and AVOs had more primary penalties than co-located teams and AVOs,

whereas co-located teams and AVOs had more secondary penalty points. Likewise, in Experiment 1, distributed DEMPCs received more primary penalties than co-located DEMPCs.

Effects of Workload on Perceived Workload

The analysis of dual task performance (i.e., primary versus secondary) is only one of several ways to examine the effect of workload on performance in our team task. In addition to dual task performance, subjective workload ratings provide a means to observe workload effects. Subjective workload was measured with a version of the NASA TLX adapted to our task. Following each mission, participants rated workload on five subscales (mental demand, physical demand, temporal demand, performance demand, and team demand). The ratings on each subscale were weighted according to the extent to which each type of demand contributes to the workload in our task (as decided by experimenters). For example, our task requires more mental demand (remembering, deciding, etc.) than physical demand (pushing, pulling, etc.) and thus, mental demand is weighted more heavily than physical demand. These weights differ among the roles, as each type of demand does not necessarily contribute to each role's workload in the same manner (see Table S18). The sum of the weighted workload subscales are divided by the sum of the weights and yields an overall workload score for each role at each mission that ranges from 0 to 100. Large numbers on the subjective ratings scale reflect higher levels of perceived workload and small numbers are indicative of lower levels of perceived workload. Team workload scores were estimated by an average of the three individual workload scores.

Table S18
The Weights for each Subscale on the NASA TLX for each Role

	Mental	Physical	Temporal	Performance	Team
AVO	1.67	1.17	2.67	2.33	2.17
PLO	2.00	.67	2.50	3.00	1.83
DEMPC	3.67	.33	1.17	1.50	3.33

The following analyses examine subjective workload at the team level as well as at each individual level using a repeated measures analysis of variance (ANOVA) with workload as the repeated factor (low vs. high) and co-located vs. distributed condition as a between-subjects factor. In the interest of capturing differences in perceived workload due to the workload manipulation, which occurred after teams reached asymptotic levels of performance (Mission 4), subjective workload ratings taken after the final low workload mission (Mission 4) were used as an estimate for low workload, and ratings taken after the first high workload mission (Mission 5) were used as an estimate for high workload. However, some graphs in this section display perceived workload across all missions for the purpose of illustrating how trends in perceived workload compare to trends in performance.

Experiment 1. Table S19 presents descriptive statistics for subjective workload at the team level. Subjective workload ratings at the team level follow inversely the trends in performance. That is, teams tended to feel less workload from Mission1 to Mission 4 (low workload) as performance increased, but experienced an increase in perceived workload after the first high workload mission (Mission 5) was introduced. Figure S1 shows the subjective workload rating at each mission at the team level, averaged across the three individuals. Figure S2 displays

overall team performance at each mission for co-located and distributed teams for the purpose of comparing the trends of performance and subjective workload. Analyses on performance are reported in an earlier section.

A significant effect of workload emerged such that there was a significant increase in the perception of workload in high workload, $F(1, 18) = 17.13$, $p < .01$. There was no main effect of dispersion, $F(1, 18) < 1$, or dispersion by workload interaction, $F(1, 18) < 1$.

Table S19
NASA TLX Ratings for Co-located and Distributed Teams during Low and High Workload in Experiment 1

	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Mission 4 (LW)	56.39	52.36	14.67	8.07	37.27	35.95	77.65	62.31
Mission 5 (HW)	63.13	59.22	14.83	12.40	39.88	37.67	83.44	79.70

LW = Low Workload
HW = High Workload

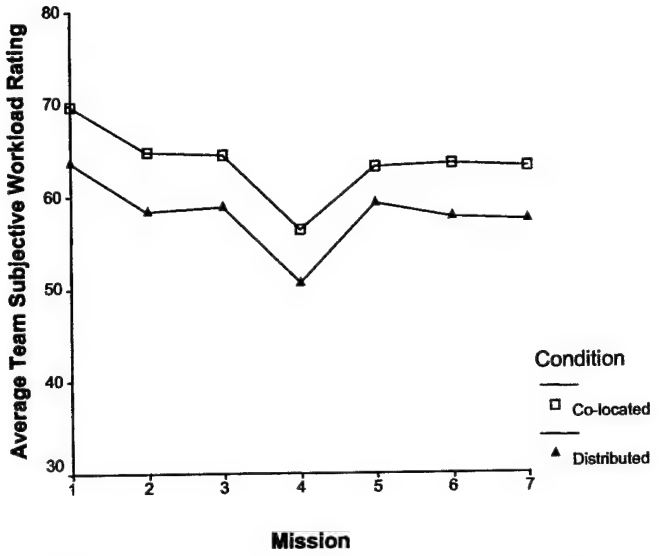


Figure S1. Average subjective workload ratings for co-located and distributed teams at each mission in Experiment 1.

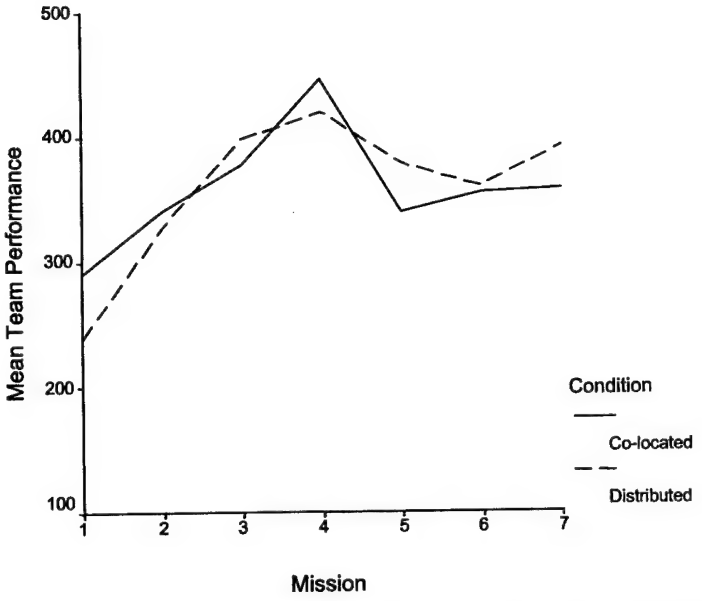


Figure S2. Average performance for co-located and distributed teams at each mission in Experiment 1.

Similar patterns in the subjective workload ratings were found at the individual levels. Table S20 presents descriptive statistics for subjective workload ratings for each of the three team roles, separately for co-located and distributed conditions. These data are graphically presented in Figure S3. Figure S4 shows the performance for each role, separately for co-located and distributed, for the purpose of comparing trends in subjective workload ratings and performance. In looking at Figure S4, it is important to keep in mind that the performance scores for each role are calculated in a unique way and thus it is not meaningful to compare performance scores across roles.

Table S20
NASA TLX Ratings for Co-located and Distributed AVOs, PLOs, and DEMPCs during Low and High Workload in Experiment 1

	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Mission 4 (LW)	56.84	50.61	13.35	18.32	36.05	24.62	72.24	78.49
Mission 5 (HW)	58.61	56.94	15.91	22.22	26.90	23.65	78.27	94.29
Mission 4 (LW)	56.51	62.49	19.91	15.84	34.05	26.08	89.71	81.06
Mission 5 (HW)	63.22	65.74	21.41	15.21	31.93	41.66	93.69	88.46
Mission 4 (LW)	55.80	42.82	23.97	18.66	19.67	14.52	87.34	64.13
Mission 5 (HW)	67.57	54.98	26.33	18.44	25.47	27.66	96.83	79.70

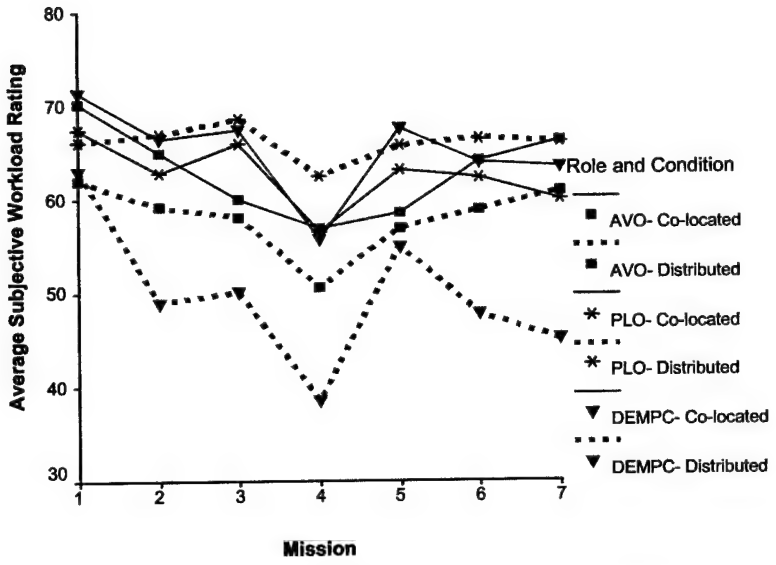


Figure S3. Average subject workload ratings for co-located and distributed AVOs, PLOs, and DEMPCs in Experiment 1.

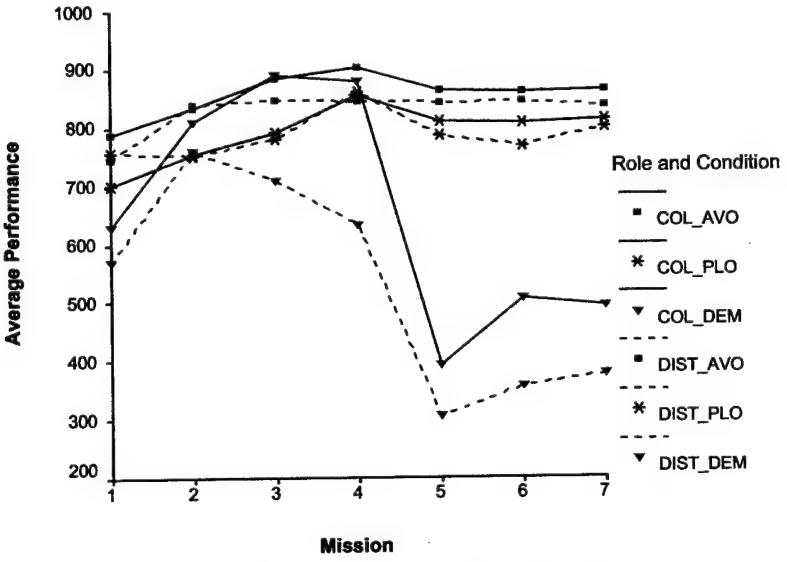


Figure S4. Average performance for co-located and distributed AVOs, PLOs, and DEMPCs in Experiment 1.

A repeated measures ANOVA was run in order to examine the effects of role (AVO, PLO, DEMPC), workload, and dispersion. There was a significant effect of workload, $F(1, 54) = 14.52$, $p < .01$, indicating that perceived levels of workload changed significantly from Mission 4 to Mission 5. Neither a main effect of dispersion, $F(1, 54) = 1.08$, nor a main effect of role, $F(2, 54) = 1.03$, emerged. There was also no significant interaction between workload and dispersion, $F(1, 54) = .23$, dispersion and role, $F(2, 54) = 1.40$, or workload, dispersion, and role, $F(2, 54) < 1$.

A significant interaction between workload and role emerged, $F(2, 54) = 2.51, p = .09$, suggesting that the workload manipulation differentially effected perceived workload for different team roles. As Figure S5 suggests and *post-hoc* tests revealed, only PLOs and DEMPCs perceived significant increases in workload during Mission 5, $F(1, 19) = 4.22, p = .05$ and $F(1, 19) = 8.31, p = .01$, respectively. AVOs's reports of perceived workload during Mission 4 and Mission 5 did not significantly differ, $F(1, 19) = 2.83$.

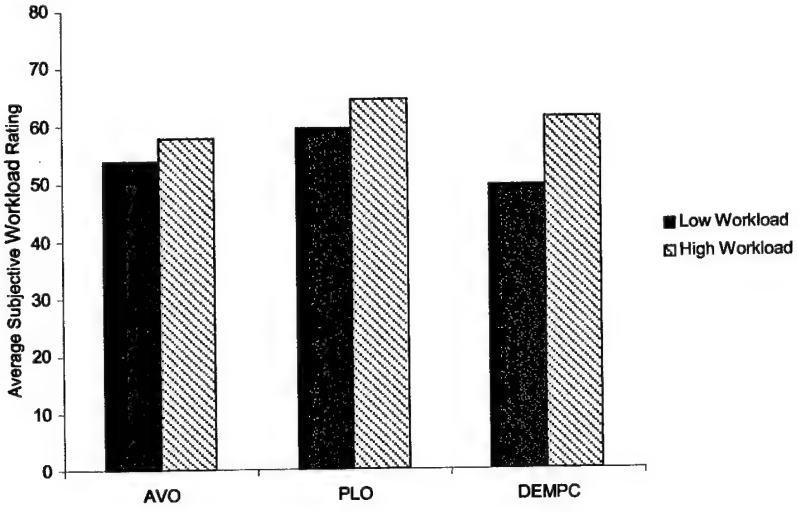


Figure S5. Average subjective workload ratings for AVOs, PLOs, and DEMPCs during low workload and high workload in Experiment 1.

The analysis above did not reveal any significant effects of dispersion on perceived workload for AVOs, PLOs, or DEMPCs despite the fact that Figure S3 depicts a considerable difference in subjective workload ratings for co-located and distributed DEMPCs. To explore this, further analyses were conducted in order to observe the effects of dispersion on DEMPCs perceived workload reported for each subscale of the NASA TLX.

An ANOVA was used to examine the difference in co-located and distributed DEMPCs' subjective workload ratings on each subscale of the NASA TLX at Mission 4 and Mission 5. The subjective ratings reported during these missions were used as the estimates for low workload and high workload, respectively. During low workload the only significant difference between co-located and distributed DEMPCs was their ratings for the performance subscale (see Table S21 for *F*-values and significance). Specifically, co-located DEMPCs perceived higher levels of workload than distributed DEMPCs concerning how they would perform.

Table S21
Descriptive and Inferential Statistics of each Subjective Workload Subscale for Co-located and Distributed DEMPCs at Mission 4 in Experiment 1

Subscale	Mean		Standard Deviation		*F*-value Col vs. Dist *df* = 1, 19
	COL	DIST	COL	DIST	
Mental	193.78	162.21	123.24	108.91	.37
Physical	3.99	3.14	5.93	3.00	.17
Temporal	58.62	44.93	35.87	32.12	.81
Performance	121.80	66.90	24.42	53.86	8.62*
Team	179.82	108.23	124.45	81.78	2.31

* $p < .01$

The difference in subjective workload ratings on the performance subscale for co-located and distributed DEMPCs in low workload was also found in high workload (see Table S22 for *F*-values and significance). No other differences in subjective ratings for co-located and distributed teams were found.

Table S22
Descriptive and Inferential Statistics of each Subjective Workload Subscale for Co-located and Distributed DEMPCs at Mission 5 in Experiment 1

Subscale	Mean		Standard Deviation		*F*-value Col vs. Dist *df* = 1, 19
	Col	Dist	Col	Dist	
Mental	271.58	236.35	122.66	100.30	.49
Physical	3.50	7.10	4.11	5.95	2.47
Temporal	77.57	65.40	39.42	33.43	.55
Performance	122.25	76.50	24.13	49.88	6.82*
Team	200.80	164.50	118.60	94.44	.57

* $p < .05$

Overall, the subjective workload ratings appeared to be consistently high for some subscales (e.g., mental and team) and consistently low for other subscales (e.g., physical). However, the significant differences found between co-located and distributed DEMPCs on the performance subscale indicate that the co-located DEMPCs felt more pressure to perform well. Interestingly, these DEMPCs were on teams who could (1) see each other and (2) see each others individual performance score following each mission.

Experiment 2. Table S23 presents descriptive statistics for subjective workload at the team level. Subjective workload ratings at the team level follow inversely the trends in performance as seen in Figures S6 and S7.

Repeated measures ANOVA revealed significant effect of workload at the team level, $F(1, 18) = 12.86$, $p < .01$, meaning that team perceived higher workload when high workload mission was introduced. There was no main effect of dispersion, $F(1, 18) < 1$, or dispersion by workload interaction, $F(1, 18) < 1$.

Table S23
NASA TLX Ratings for Co-located and Distributed Teams during Low and High Workload in Experiment 2

	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Mission 4 (LW)	61.85	61.91	15.82	12.78	30.85	41.84	92.50	94.02
Mission 5 (HW)	62.47	60.75	18.57	20.35	18.26	26.65	98.10	97.74

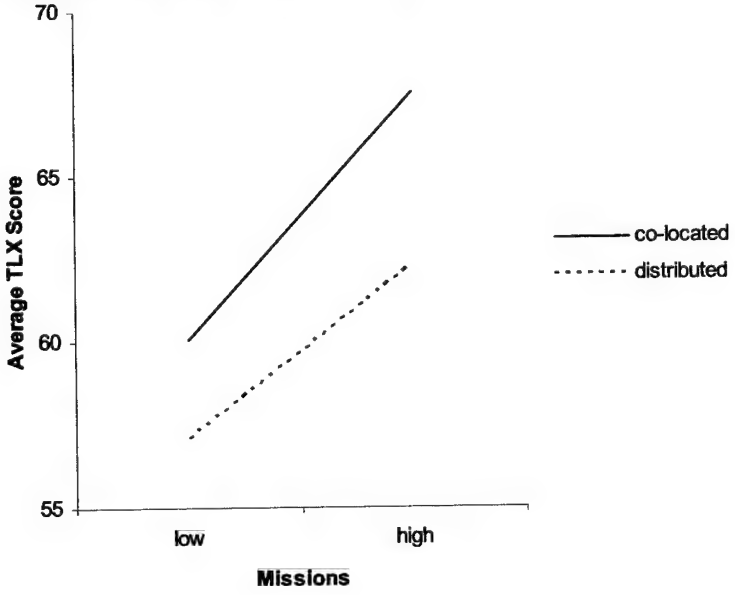


Figure S6. Average subjective workload ratings for co-located and distributed teams at Missions 4 and 5.

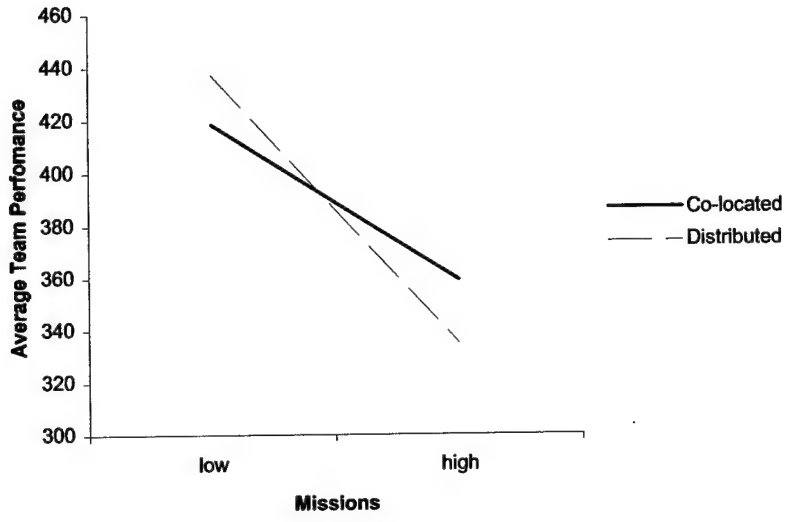


Figure S7. Average performance for co-located and distributed teams at missions 4 and 5.

A 2 (dispersion) by 2 (workload) by 3 (role) ANOVA was run in order to examine the relationship between role and subjective workload rating following each mission, for each dispersion condition. Dispersion and role were between-subjects factors and workload was a within-subjects factor. A main effect of dispersion was not detected, $F(1, 54) = 1.16$. Also there was no significant interaction between workload and dispersion $F(1, 54) < 1$, workload and role $F(1, 54) = 1.74$, or workload, role, and dispersion $F(1, 54) < 1$. However, this analysis revealed a significant effect of workload $F(1, 54) = 16.27$, $p < .01$, indicating that perceived levels of workload increased significantly between missions. Also main effect of role was significant $F(2, 54) = 2.61$, $p = .08$. Further, there was a significant interaction between role and dispersion $F(2, 54) = 2.94$, $p = .06$. In Table S24 we see that co-located and distributed DEMPCs reported significantly different subjective workload ratings $F(1, 18) = 9.66$, $p = .06$, while co-located and distributed AVOs and PLOs tended to give more similar responses.

Table S24
NASA TLX Ratings for Co-located and Distributed AVOs, PLOs, and DEMPCs during Low and High Workload in Experiment 2

	Mean		Standard Deviation		Minimum		Maximum	
	Col	Dist	Col	Dist	Col	Dist	Col	Dist
Mission 4 (LW)	59.35	62.85	15.88	11.62	41.47	44.47	89.82	77.00
Mission 5 (HW)	65.64	65.97	16.57	11.10	49.27	51.38	94.02	80.90
Mission 4 (LW)	51.44	56.34	11.87	17.75	30.85	27.17	71.94	84.43
Mission 5 (HW)	57.21	58.15	9.93	22.18	42.43	26.46	77.24	97.74
Mission 4 (LW)	69.44	52.15	17.66	16.87	45.67	18.26	96.70	77.36
Mission 5 (HW)	79.62	62.74	17.88	16.75	44.11	34.96	98.00	92.50

LW = Low Workload
HW = High Workload

Due to the significant difference in co-located and distributed DEMPC's subjective ratings, further analyses were conducted in order to observe the effects of dispersion on DEMPC's perceived workload reported for each subscale of the questionnaire. During low workload the only significant difference between co-located and distributed DEMPCs was in their ratings on the team demands subscale (see Table S25 for *t*-values and significance). Specifically, co-located DEMPCs perceived higher levels of team demands than distributed DEMPCs, who were physically separated from the other team members.

Table S25
Descriptive and Inferential Statistics of each Subjective Workload Subscale for Co-located and Distributed DEMPCs at Mission 4 in Experiment 2

Subscale	Mean		Standard Deviation		T-value Col vs. Dist $df = 1, 19$
	COL	DIST	COL	DIST	
Mental	265.34	255.07	82.16	83.32	.26
Physical	3.40	6.24	3.61	9.85	-.78
Temporal	63.06	45.63	31.16	26.18	1.16
Performance	121.20	112.35	32.22	50.53	.41
Team	241.43	102.23	115.34	105.61	2.70*

*$p < .05$

The difference in the subscale ratings for co-located and distributed DEMPCs found in the low workload mission was also found in the high workload mission (see Table S26). Moreover, a significant difference on the temporal demand subscale also emerged. Apparently, during high workload the distributed DEMPCs experience less temporal demands compared to co-located DEMPCs.

Table S26
Descriptive and Inferential Statistics of each Subjective Workload Subscale for Co-located and Distributed DEMPCs at Mission 5 in Experiment 2

Subscale	Mean		Standard Deviation		T-value Col vs. Dist $df = 1, 19$
	COL	DIST	COL	DIST	
Mental	304.61	286.63	74.33	75.00	.71
Physical	6.60	6.77	7.01	10.58	-.04
Temporal	98.51	71.60	20.35	25.10	2.83*
Performance	129.75	123.90	24.62	34.43	.37
Team	256.74	138.53	102.00	109.40	3.07**

*$p \leq .05$ **$p \leq .01$

Summary. Results from the NASA TLX in both experiments indicated that perceived workload depended on roles and conditions. In both experiments co-located DEMPCs perceived greater workload demands (either performance, team, or temporal) than distributed DEMPCs. AVOs in Experiment 1 perceived no change in workload, but PLOs did. In Experiment 2 distributed AVOs and PLOs perceived greater workload than distributed DEMPCs.

Comparison of Various Measures of Workload

Analyses of the workload's effects on dual task performance as well as the analyses of perceived workload have shown that, in some form, workload influenced teams' ability to perform in our UAV task. In some cases, primary task performance suffered more than secondary task performance in high workload. Moreover, the co-located versus distributed condition also determined how workload influenced performance. Furthermore, some team positions perceived more workload than others and felt more pressure to do well. This section examines how well these measures of workload effects converge.

The difference between the low and high workload primary task performance penalties, secondary task performance penalties, and the TLX measure were subjected to correlational analyses. Table S27 shows the correlations among the three dependent measures in Experiment 1. A significant correlation was found between secondary task performance penalty and the subjective ratings indicating that teams with large differences between the amount of secondary task performance penalties received in Mission 4 and Mission 5 also tended to rate their perception of workload more differently from Mission 4 to Mission 5. No significant correlations were found in Experiment 2 (Table S28).

Table S27
Correlations among Workload Measures during Low Workload and High Workload in Experiment 1

	Difference Between LW and HW Primary Task Performance Penalty	Difference Between LW and HW Secondary Task Performance Penalty
Difference Between LW and HW Secondary Task Performance Penalty	.16	
Difference Between LW and HW Subjective Ratings	-.09	.39*

*$p < .10$

Table S28
Correlations among Workload Measures during Low Workload and High Workload in Experiment 2

	Difference Between LW and HW Primary Task Performance Penalty	Difference Between LW and HW Secondary Task Performance Penalty
Difference Between LW and HW Secondary Task Performance Penalty	.13	
Difference Between LW and HW Subjective Ratings	-.21	.01

Summary. As the correlations show, these performance-based measures and subjective measure of workload effects are not consistently linearly related. However, the subjective and objective measures of workload did converge on some effects. For example, Experiment 1 DEMPCs, whose primary performance suffered in high workload, also found the task to be more demanding in high workload.

The analyses have revealed some of the complexity of the team workload construct. These results have also demonstrated the importance of examining the effects of workload at the team level as well as at the role level. The unique nature of each of the three roles is a primary characteristic of our UAV-STE task. These measures of workload are consistent in that they both highlight differences in individual and team workload. For instance, in Experiment 1 at the team level, primary task performance deteriorates during high workload and teams in general seem to perceive this increase in workload. However, this is only true for two of the three roles (PLOs and DEMPCs). Likewise, although team performance in Experiment 1 is affected by workload, this is not the case for each role individually (only the DEMPC is affected). Through comparisons of the effects of workload at the individual and team levels, we can begin to understand the impact of each role on the team.

Appendix T

Proportion of Agreement Index for Process Measures in Experiment 1

Question	Mission 1				Mission 2			
	t	*df*	*p*	Mean Agreement	*t*	*df*	*p*	Mean Agreement
P1	15.77	19	0.000	0.800	9.75	19	0.000	0.667
P2	11.46	19	0.000	0.775	2.87	19	0.010	0.650
P3	7.55	19	0.000	0.750	6.66	19	0.000	0.700
P4	7.55	19	0.000	0.750	8.72	19	0.000	0.800
P5	10.72	19	0.000	0.725	7.43	19	0.000	0.675
P6	19.00	19	0.000	0.950	10.38	19	0.000	0.850
Comm/Crd	17.33	19	0.000	0.788	22.34	19	0.000	0.850
Decision Mkg	22.34	19	0.000	0.850	30.23	19	0.000	0.863
SA Behaviors	28.79	19	0.000	0.788	20.68	19	0.000	0.750
Overall	25.52	19	0.000	0.838	30.49	19	0.000	0.938

Question	Mission 3				Mission 4			
	t	*df*	*p*	Mean Agreement	*t*	*df*	*p*	Mean Agreement
P1	12.84	19	0.000	0.767	10.42	19	0.000	0.667
P2	16.91	19	0.000	0.925	16.91	19	0.000	0.925
P3	6.66	19	0.000	0.700	4.82	19	0.000	0.550
P4	13.08	19	0.000	0.900	7.55	19	0.000	0.750
P5	11.00	19	0.000	0.825	11.57	19	0.000	0.850
P6	13.08	19	0.000	0.900	13.08	19	0.000	0.900
Comm/Crd	15.08	19	0.000	0.825	21.00	19	0.000	0.788
Decision Mkg	16.34	19	0.000	0.863	22.33	19	0.000	0.838
SA Behaviors	18.42	19	0.000	0.825	22.76	19	0.000	0.813
Overall	16.81	19	0.000	0.888	27.36	19	0.000	0.800

Question	Mission 5				Mission 6			
	t	*df*	*p*	Mean Agreement	*t*	*df*	*p*	Mean Agreement
P1	11.05	19	0.000	0.750	14.69	19	0.000	0.833
P2	8.82	19	0.000	0.750	7.85	19	0.000	0.725
P3	7.55	19	0.000	0.750	4.36	19	0.000	0.500
P4	8.72	19	0.000	0.800	8.72	19	0.000	0.800
P5	5.64	19	0.000	0.600	3.58	19	0.002	0.325
P6	10.38	19	0.000	0.850	-	-	-	1.000
Comm/Crd	15.20	19	0.000	0.775	30.25	19	0.000	0.850
Decision Mkg	14.31	19	0.000	0.825	23.13	19	0.000	0.888
SA Behaviors	20.34	19	0.000	0.700	25.84	19	0.000	0.825
Overall	15.16	19	0.000	0.838	30.25	19	0.000	0.850

Question	Mission 7			
	t	*df*	*p*	Mean Agreement
P1	11.41	19	0.000	0.800
P2	8.11	19	0.000	0.750
P3	6.66	19	0.000	0.700
P4	10.38	19	0.000	0.850
P5	5.51	19	0.000	0.575
P6	10.38	19	0.000	0.850
Comm/Crd	22.48	19	0.000	0.863
Decision Mkg	25.42	19	0.000	0.850
SA Behaviors	20.29	19	0.000	0.813
Overall	32.03	19	0.000	0.900

Appendix U

Proportion of Agreement Index for Process Measures in Experiment 2

Question	Mission 1				Mission 2			
	t	*df*	*p*	Mean Agreement	*t*	*df*	*p*	Mean Agreement
P1	14.26	18	0.000	0.842	15.09	18	0.000	0.789
P2	15.37	18	0.000	0.842	-	-	-	1.000
P3	3.92	11	0.002	0.583	6.20	14	0.000	0.733
P4	6.71	15	0.000	0.750	6.90	13	0.000	0.786
P5	11.45	13	0.000	0.786	14.67	14	0.000	0.867
P6	10.25	15	0.000	0.875	7.42	11	0.000	0.833
Comm/Crd	19.29	18	0.000	0.829	22.80	16	0.000	0.838
Decision Mkg	29.52	18	0.000	0.868	23.10	16	0.000	0.824
SA Behaviors	24.82	18	0.000	0.829	27.73	16	0.000	0.853
Overall	30.75	18	0.000	0.895	23.31	16	0.000	0.882

Question	Mission 3				Mission 4			
	t	*df*	*p*	Mean Agreement	*t*	*df*	*p*	Mean Agreement
P1	8.92	15	0.000	0.729	13.79	16	0.000	0.765
P2	31.00	15	0.000	0.969	23.37	16	0.000	0.941
P3	5.74	15	0.000	0.688	5.42	16	0.000	0.647
P4	7.48	14	0.000	0.800	6.71	15	0.000	0.750
P5	12.36	13	0.000	0.821	21.96	15	0.000	0.938
P6	5.74	11	0.000	0.750	8.83	13	0.000	0.857
Comm/Crd	16.78	15	0.000	0.813	23.31	16	0.000	0.882
Decision Mkg	20.19	15	0.000	0.781	28.06	16	0.000	0.838
SA Behaviors	18.82	15	0.000	0.828	28.06	16	0.000	0.838
Overall	21.80	15	0.000	0.844	28.28	16	0.000	0.882

Question	Mission 5			
	t	*df*	*p*	Mean Agreement
P1	16.66	17	0.000	0.778
P2	17.63	17	0.000	0.889
P3	8.06	15	0.000	0.813
P4	7.48	14	0.000	0.800
P5	21.00	10	0.000	0.955
P6	10.95	16	0.000	0.882
Comm/Crd	19.51	17	0.000	0.778
Decision Mkg	24.50	17	0.000	0.889
SA Behaviors	19.00	17	0.000	0.792
Overall	28.66	17	0.000	0.847